Qawā'id
Jurnal Bahasa dan Sastra Arab
Vol. 2 No. 03 Tahun 2026 [Page 158-166]
E-ISSN 3123-5484

تحليل جودة أسئلة الاختبار النهائي في اللغة العربية للصف الثاني المتوسط بمدرسة لقمان الحكيم

Lailan Fathiyah
Sekolah Tinggi Agama Islam As-Sunnah Deli Serdang, Indonesia
Email: lailanfathiyah@gmail.com

Abstract

This study aims to determine the quality of the items in the end-of-semester Arabic language examination for eighth-grade students at MTs Luqman Al-Hakim Islamiyah Deli Serdang, based on validity, reliability, difficulty level, discrimination power, and distractor effectiveness. The study employed a descriptive quantitative approach. Data were collected through documentation, consisting of the 25 multiple-choice test items, the answer sheets of 37 female students, the answer key, the test blueprint, and the lesson plan (RPP), and were analyzed using ANATES version 4.0.2. The results show that the correlation coefficient (rXY) was 0.77, placing the test in the good validity category, while content validity analysis found that 24 items (96%) were valid and one item (4%) was invalid because it did not match the blueprint indicator. The reliability coefficient was 0.87, categorized as very high. The difficulty distribution was unbalanced, with 2 difficult items (8%), 12 moderate items (48%), and 11 easy items (44%). In terms of discrimination power, 3 items (12%) were excellent, 12 (48%) good, 9 (36%) fair, and 1 (4%) poor, with no negative values. Distractor effectiveness was rated excellent for 8% of the items, good for 36%, moderate for 36%, and poor for 20%. The study concludes that the test is generally good in terms of validity, reliability, and discrimination power, but needs improvement in the distribution of difficulty levels and the quality of distractors.

Keywords: *validity, reliability, difficulty level, discrimination power, distractor effectivene*

المستخلص

هدفت هذه الدراسة إلى معرفة جودة بنود أسئلة الاختبار النهائي للفصل الدراسي لمادة اللغة العربية للصف الثاني المتوسط بمدرسة لقمان الحكيم الإسلامية ديلي سردانج من حيث الصدق والثبات ومستوى الصعوبة وقوة التمييز وفعالية المشتتات .استخدمت الدراسة المنهج الكمي الوصفي، وجُمعت البيانات بالتوثيق من ورقة الأسئلة)25 بنداً من نوع الاختيار من متعدد(وأوراق إجابات سبع وثلاثين طالبة ومفتاح الإجابات وجدول المواصفات وخطة تنفيذ التعليم، وحُلّلت بتطبيق ANATES الإصدار 4.0.2 .أظهرت النتائج أن معامل الارتباط)الصدق(بلغ 0,77 في فئة جيدة، وأن 24 بنداً)96%(صحيحة من حيث صدق المحتوى وبنداً واحداً)4%(غير صحيح لعدم توافقه مع جدول المواصفات، وأن معامل الثبات بلغ 0,87 في فئة عالية جداً%، وكان توزيع مستويات الصعوبة غير متوازن)بندان صعبان بنسبة 8%، و12 بنداً متوسطاً بنسبة 48 %، و11 بنداً سهلاً بنسبة 44(% .وفي قوة التمييز جاء 3 بنود)12%(ممتازة و12)48%(جيدة و9)36%(مقبولة وبند واحد)4%(ضعيف، دون أي قيمة سالبة .أما المشتتات فكانت 8 %ممتازة و36 %جيدة و36 %متوسطة و20 %ضعيفة .وتخلص الدراسة إلى أن الاختبار جيد عموماً من حيث الصدق والثبات وقوة التمييز، لكنه يحتاج إلى تحسين في توزيع مستويات الصعوبة وجودة المشتتات.

الكلمات المفتاحية: الصدق، الثبات، مستوى الصعوبة، قوة التمييز، فعالية المشتتات

المقدمة

يُعدّ التقويم عنصراً أساسياً في العملية التعليمية، لما يؤديه من دور في الكشف عن مدى تحقق الأهداف التعليمية لدى المتعلمين، وتزويد المعلم بالمعلومات اللازمة لاتخاذ القرارات المناسبة بشأن تحسين عملية التعليم والتعلّم وتطويرها .(Arifin, 2017) ويُعدّ الاختبار من أكثر أدوات التقويم استخداماً في المؤسسات التعليمية، ولا سيما الاختبار النهائي للفصل الدراسي، الذي يُستخدم لقياس مستوى التحصيل الدراسي لدى الطلاب في نهاية فترة تعليمية محددة، وتحديد مدى استيعابهم للمادة الدراسية التي تلقّوها (Hermawan, 2021). وفي هذا السياق، يحظى اختبار اللغة العربية بأهمية خاصة، نظراً إلى ارتباطه بقياس مدى اكتساب المتعلمين للمعارف والمهارات اللغوية المستهدفة. ومن ثمّ، فإن جودة أسئلة الاختبار تمثل أحد العوامل المهمة التي تؤثر في دقة نتائج التقويم ومدى إمكان الاعتماد عليها في الحكم على مستوى التحصيل الدراسي للطلاب (بيهقي et al., 2025).

غير أن أهمية الاختبار لا تتحدد بمجرد استخدامه في قياس التحصيل الدراسي، بل ترتبط ارتباطاً وثيقاً بجودة بنوده وخصائصها القياسية. فالاختبار الذي لا يقيس ما وُضع لقياسه، أو الذي لا تتسم نتائجه بالاتساق، أو الذي يعجز عن التمييز بين الطلاب ذوي المستويات التحصيلية المختلفة، قد يؤدي إلى نتائج مضللة وأحكام غير دقيقة بشأن كفاءة المتعلمين. وقد يترتب على ذلك اتخاذ قرارات تعليمية لا تعكس المستوى الحقيقي للطلاب، فضلاً عن صعوبة تحديد جوانب القوة والضعف في تعلمهم. لذلك، لا يكفي إعداد الأسئلة وتطبيقها وتصحيحها، بل ينبغي إخضاعها لتحليل علمي منظّم للكشف عن مدى جودتها، وتحديد البنود التي يمكن استخدامها، وتلك التي تحتاج إلى تعديل أو استبدال. وفي هذا الإطار، أشارت دراسات سابقة إلى أهمية تحليل بنود الاختبار في تحسين جودة الأسئلة وتعزيز كفاءة المعلم في إعداد أدوات التقويم (Savika & Zuhriyah, 2024).

ويُعدّ تحليل البنود في ضوء النظرية الكلاسيكية للاختبارات (Classical Test Theory) من المداخل المستخدمة لتقييم جودة الاختبارات وتحليل خصائصها القياسية .(Sumaryanta, 2021) ويستند هذا التحليل إلى خمسة مؤشرات رئيسة، هي الصدق، والثبات، ومستوى الصعوبة، وقوة التمييز، وفعالية المشتتات. ويشير الصدق إلى مدى قدرة الاختبار على قياس السمة أو التحصيل الذي أُعدّ لقياسه(Asyim et al., 2024) ، في حين يشير الثبات إلى مدى اتساق نتائج الاختبار وإمكان الاعتماد عليها عند تفسير أداء المتعلمين (Farida & Musyarofah, 2021؛ حنان وبوعموشة، 2020). أما مستوى الصعوبة، فيعبّر عن نسبة الطلاب الذين أجابوا عن البند إجابة صحيحة، ويُستخدم لتحديد ما إذا كان البند سهلاً أو متوسط الصعوبة أو صعباً (Arifin, 2017). وتشير قوة التمييز إلى قدرة البند على التفريق بين الطلاب ذوي التحصيل المرتفع والطلاب ذوي التحصيل المنخفض، بينما تعبّر فعالية المشتتات عن مدى قدرة البدائل الخاطئة في أسئلة الاختيار من متعدد على جذب الطلاب الذين لم يتمكنوا من تحديد الإجابة الصحيحة .(Handriawan & Nurman, 2021) وتتيح هذه المؤشرات مجتمعةً فهماً أكثر شمولاً لجودة الاختبار، إذ لا يقتصر التقييم على مدى ملاءمة الأسئلة من حيث المحتوى، بل يشمل أيضاً خصائصها الإحصائية ومدى فاعليتها في قياس الفروق بين مستويات أداء المتعلمين.

وفي ضوء التطور في استخدام الأدوات الحاسوبية في مجال التقويم التربوي، أصبح برنامج ANATESمن الأدوات المستخدمة في تحليل بنود الاختبارات، لما يتيحه من إمكانات في حساب المؤشرات الإحصائية اللازمة لتقييم جودة الأسئلة .(Aini et al., 2025) وقد استُخدم هذا البرنامج في دراسات تناولت تحليل أسئلة اللغة العربية، مما يعكس إمكان الاستفادة منه في فحص خصائص بنود الاختبارات اللغوية .(Dianova & Anwar, 2024)كما استُخدمت النظرية الكلاسيكية للاختبارات في تحليل جودة أسئلة الاختيار من متعدد(Nasyrah et al., 2024) ، وكذلك في تحليل أدوات اختبارية أخرى ضمن سياقات تعليمية مختلفة .(Warju et al., 2020) وتبرز أهمية هذا المدخل في أنه يساعد المعلمين والباحثين على اتخاذ قرارات

مبنية على بيانات فعلية بشأن صلاحية البنود، بدلاً من الاعتماد على الانطباعات الشخصية وحدها عند تقييم جودة الأسئلة.

وقد تناولت دراسات سابقة تحليل جودة أسئلة اللغة العربية باستخدام برنامجANATES ، وقدمت نتائج تسهم في توضيح واقع جودة البنود الاختبارية في عدد من المؤسسات التعليمية. فقد أجرى Nurlaeli (2024)دراسة لتحليل أسئلة التقويم التجميعي النصفي للصف الثامن في مدرسة MTs Ma'arif NU 1 Patikraja Banyumas، وأظهرت النتائج أن معظم البنود كانت متوسطة الصعوبة، في حين لم تتجاوز نسبة البنود الصالحة للاستخدام دون مراجعة 20%، بينما احتاجت نسبة 70% من البنود إلى التعديل، وصُنّفت نسبة 10% منها على أنها غير صالحة. وتشير هذه النتائج إلى أن ملاءمة مستوى صعوبة الأسئلة لا تعني بالضرورة ارتفاع جودتها من جميع الجوانب، وأن الحكم على جودة الاختبار يتطلب النظر في مجموعة من المؤشرات المتكاملة.

وفي دراسة أخرى، حلّلت Febriyanti (2024) أسئلة الاختبار النهائي لمادة اللغة العربية للصف الأول المتوسط في مدرسة الجمعية الوصلية 40 تينوكّة، وأظهرت النتائج انخفاض مستوى الصدق والثبات، الأمر الذي استدعى مراجعة عدد كبير من الأسئلة. وتؤكد هذه النتيجة أهمية التحقق من الخصائص القياسية للاختبار قبل الاعتماد على نتائجه في تقويم التحصيل الدراسي. أما(2024) Roudhana ، فقد تناولت أسئلة اللغة العربية في معهد دار الحكمة، وتوصلت إلى أن غالبية البنود اتسمت بمستوى متوسط من الصعوبة، وتمييز جيد نسبياً، ومشتتات فعالة، مع وجود بعض البنود التي لا تزال بحاجة إلى المراجعة والتحسين. وتوضح هذه الدراسات أن جودة أسئلة اللغة العربية قد تختلف من مؤسسة تعليمية إلى أخرى، وأن تحليل البنود يوفر أساساً علمياً لتحديد جوانب القوة والقصور في أدوات التقويم.

وعلى الرغم من إسهام الدراسات السابقة في توضيح خصائص أسئلة اللغة العربية في عدد من المؤسسات التعليمية، فإن نتائجها تشير إلى أهمية مواصلة البحث في هذا المجال ضمن سياقات تعليمية مختلفة، مع مراعاة طبيعة الاختبار والفئة المستهدفة ومجموعة المؤشرات المستخدمة في التحليل. وتتمثل الفجوة البحثية التي تنطلق منها الدراسة الحالية في الحاجة إلى تقديم تحليل متكامل لجودة أسئلة الاختبار النهائي للفصل الدراسي في مادة اللغة العربية لدى طالبات الصف الثاني المتوسط بمدرسة لقمان الحكيم الإسلامية ديلي سردانج، من خلال الجمع بين مؤشرات الصدق والثبات ومستوى الصعوبة وقوة التمييز وفعالية المشتتات في إطار تحليلي واحد. وتنبع أهمية هذا التوجه من أن بعض الدراسات السابقة ركزت على جوانب معينة من جودة البنود أو أبرزت نتائج مؤشرات محددة، في حين تسعى الدراسة الحالية إلى فحص المؤشرات الخمسة مجتمعةً، بما يسمح بتكوين صورة أكثر شمولاً عن جودة الاختبار وملاءمة بنوده للاستخدام في التقويم التعليمي.

وتتجلى الجدة العلمية للدراسة الحالية في ثلاثة جوانب مترابطة. أولها، اعتماد منظور تحليلي تكاملي يجمع المؤشرات الخمسة لجودة الاختبار بدلاً من الاكتفاء بمؤشر واحد أو مجموعة محدودة من المؤشرات. وثانيها، توجيه التحليل إلى أسئلة الاختبار النهائي للفصل الدراسي في مادة اللغة العربية لدى طالبات الصف الثاني المتوسط في مدرسة لقمان الحكيم، بما يضيف نتائج ميدانية مرتبطة بسياق تعليمي محدد لم تتناوله الدراسات السابقة المذكورة. وثالثها، توظيف نتائج التحليل لتحديد جوانب القوة والقصور في بنود الاختبار، وتوفير أساس علمي يمكن الاستفادة منه في اتخاذ قرارات تتعلق بالإبقاء على البنود أو تعديلها أو استبعادها. وبذلك، لا تقتصر الدراسة على وصف الخصائص الإحصائية للأسئلة، بل تهدف أيضاً إلى الإسهام في تحسين جودة أدوات تقويم التحصيل الدراسي في مادة اللغة العربية.

وانطلاقاً من هذه الخلفية، تهدف الدراسة الحالية إلى تحليل جودة أسئلة الاختبار النهائي للفصل الدراسي في مادة اللغة العربية للصف الثاني المتوسط بمدرسة لقمان الحكيم، بالاستناد إلى مؤشرات الصدق والثبات ومستوى الصعوبة وقوة التمييز وفعالية المشتتات. وتسعى الدراسة إلى الإجابة عن السؤال الرئيس الآتي: ما مدى جودة أسئلة الاختبار النهائي للفصل الدراسي في مادة اللغة العربية للصف الثاني المتوسط

بمدرسة لقمان الحكيم في ضوء مؤشرات الصدق والثبات ومستوى الصعوبة وقوة التمييز وفعالية المشتتات؟ ومن المتوقع أن تسهم نتائجها في توفير معلومات علمية تساعد على تحسين إعداد الاختبارات وتطوير ممارسات التقويم في تعليم اللغة العربية.

منهج البحث

اعتمدت الدراسة المنهج الكمي الوصفي، لملاءمته لطبيعة البحث الذي يهدف إلى وصف جودة بنود الاختبار النهائي في مادة اللغة العربية وتحليل خصائصها استناداً إلى البيانات العددية والقيم الإحصائية، دون اختبار الفروض البحثية .(Sudaryono, 2019) وأُجريت الدراسة في مدرسة لقمان الحكيم الإسلامية ديلي سردانج، بمقاطعة سومطرة الشمالية، خلال الفترة الممتدة من سبتمبر 2025م إلى مايو 2026م. وتكوّن مجتمع الدراسة وعينتها من خمسة وعشرين (25) بنداً من بنود الاختبار النهائي، إضافة إلى أوراق إجابات سبع وثلاثين (37) طالبة من طالبات الصف الثاني المتوسط. واعتمدت الدراسة أسلوب العينة الشاملة، نظراً إلى محدودية عدد مفردات المجتمع، إذ جرى تحليل جميع البنود وأوراق الإجابات المتاحة دون اختيار عينة جزئية، استناداً إلى ما أشار إليه أريكونتو كما ورد في .(Mushofa et al., 2024) وقد أتاح هذا الإجراء فحص جميع بنود الاختبار في ضوء المؤشرات المحددة، وتوفير بيانات كمية مناسبة لتقييم جودتها بصورة منهجية.

وجُمعت البيانات باستخدام طريقة التوثيق(Sugiyono, 2023) ، من خلال مراجعة مجموعة من الوثائق ذات الصلة، شملت ورقة أسئلة الاختبار النهائي، وأوراق إجابات الطالبات، ومفتاح الإجابات، وجدول مواصفات الأسئلة، وخطة تنفيذ التعليم .(RPP) وبعد ذلك، أُدخلت البيانات في برنامج ANATES الإصدار 4.0.2، بدءاً بتحديد عدد المفحوصين والبنود والخيارات لكل بند، ثم إدخال مفتاح الإجابات، وإجراء المعالجة الإحصائية باستخدام البرنامج لاستخراج مؤشرات الصدق والثبات ومستوى الصعوبة ومعامل التمييز وفعالية المشتتات .(Sihombing et al., 2024) ويُسهم استخدام البرنامج في تسريع عملية التحليل وتيسير إجراء الحسابات الإحصائية والحد من الأخطاء البشرية المحتملة، إلا أن موثوقية النتائج تظل مرتبطة بصحة البيانات المدخلة ودقة معالجتها؛ لذلك، حرصت الباحثة على مراجعة البيانات المدخلة والتحقق من مطابقتها للوثائق الأصلية قبل اعتماد النتائج النهائية.(Wiguna, 2021)

النتائج والمناقشة

صدق البنود وثبات الاختبار: مدى صلاحية الأداة واتساق القياس

أظهرت نتائج تحليل بنود الاختبار النهائي للفصل الدراسي في مادة اللغة العربية للصف الثاني المتوسط بمدرسة لقمان الحكيم الإسلامية ديلي سردانج أن معامل الارتباط (rXY) لأسئلة الاختيار من متعدد بلغ 0,77، وهو يقع ضمن الفاصل 0,61–0,80، ويُصنّف وفق المعيار المستخدم في الدراسة ضمن فئة «جيد». وتشير هذه النتيجة إلى وجود علاقة إيجابية جيدة بين أداء الطالبات على البنود والأداء الذي يعكسه الاختبار، مما يدعم صلاحية استخدام الأداة في تقويم التحصيل الدراسي، وفقاً لطريقة حساب معامل الارتباط المعتمدة. ومع ذلك، ينبغي تفسير هذه القيمة في ضوء نوع الصدق الذي يقيسه المعامل وإجراءات حسابه، لأن الارتباط الإحصائي لا يكفي وحده لإثبات جميع جوانب صدق الاختبار.

أما من حيث صدق المحتوى، فقد تبيّن أن 24 بنداً من أصل 25 بنداً (96%) جاءت متوافقة مع المؤشرات المحددة في جدول مواصفات الأسئلة، في حين لم يتوافق بند واحد (4%) مع المؤشر المطلوب. وتكشف هذه النتيجة أن إعداد الاختبار اتسم بدرجة مرتفعة من الاتساق مع المواصفات المحددة، إلا أن وجود بند واحد غير متوافق يوضح أن جودة الاختبار لا تتوقف على سلامة معظم البنود، بل تتطلب التحقق من ملاءمة كل بند للهدف التعليمي الذي وُضع لقياسه. وقد يؤدي عدم التوافق بين السؤال والمؤشر إلى قياس

جانب معرفي مختلف عن الجانب المستهدف، مما يضعف دقة تفسير استجابة الطالبة لذلك السؤال. ولذلك، ينبغي مراجعة البند المذكور في ضوء أهداف التعلم ومؤشرات جدول المواصفات قبل إعادة استخدامه.

وفيما يتعلق بثبات الاختبار، بلغ معامل الثبات 0,87، وهي قيمة مرتفعة تقع ضمن فئة «عالية جداً» وفق التصنيف المعتمد في الدراسة، وتتجاوز الحد الأدنى الشائع للثبات المقبول، وهو 0,70. وتدل هذه النتيجة على أن الاختبار يتمتع بدرجة مرتفعة من الاتساق الداخلي وفقاً لمعامل الثبات المستخدم، الأمر الذي يعزز إمكانية الاعتماد على نتائجه في تقويم تحصيل الطالبات. غير أن الثبات المرتفع لا يعني بالضرورة أن جميع البنود صادقة أو مناسبة؛ فقد يكون الاختبار متسقاً في قياس الأداء، مع احتوائه على بنود لا تتوافق تماماً مع الأهداف التعليمية المحددة.

ويمكن تفسير هذه النتائج في ضوء النظرية الكلاسيكية للاختبارات(Classical Test Theory) ، التي تنظر إلى الدرجة المشاهدة بوصفها حصيلة الدرجة الحقيقية والخطأ في القياس. ومن هذا المنظور، يسهم ارتفاع معامل الثبات في تقليل القلق بشأن مقدار التباين الناتج عن أخطاء القياس العشوائية، بينما يسهم تحليل الصدق في التحقق من مدى ملاءمة تفسير الدرجات للغرض الذي أُعدّ الاختبار من أجله. وبناءً على ذلك، فإن اجتماع معامل صدق قدره 0,77 ومعامل ثبات قدره 0,87 يمثل نقطة قوة في الأداة، لكنه لا يلغي الحاجة إلى مراجعة توافق البنود مع جدول المواصفات. وتنسجم هذه النتيجة مع أهمية تحليل خصائص الاختبار بصورة متكاملة، بدلاً من الاعتماد على مؤشر واحد للحكم على جودته.

مستوى الصعوبة وقوة التمييز: مدى قدرة **الاختبار على قياس الفروق بين الطالبات**

أظهر تحليل مستوى الصعوبة أن البنود الخمسة والعشرين توزعت على ثلاثة مستويات؛ إذ بلغ عدد البنود الصعبة بندين (8%)، والبنود متوسطة الصعوبة اثني عشر بنداً (48%)، والبنود السهلة أحد عشر بنداً (44%). وتوضح هذه النتائج أن البنود متوسطة الصعوبة تمثل النسبة الأكبر، لكنها لا تتجاوز نصف مجموع البنود، في حين تشكل البنود السهلة نسبة مرتفعة تبلغ 44%، مقابل نسبة محدودة من البنود الصعبة لا تتجاوز 8%. ويشير هذا التوزيع إلى ميل الاختبار نحو قياس مستويات الأداء الأساسية والمتوسطة، مع محدودية نسبية في تمثيل الأسئلة التي تتطلب مستوى أعلى من الإتقان.

وفي ضوء النظرية الكلاسيكية للاختبارات، يُعبّر مستوى الصعوبة في أسئلة الاختيار من متعدد عادةً عن نسبة المفحوصين الذين أجابوا عن البند إجابة صحيحة؛ ومن ثم، فإن ارتفاع نسبة البنود السهلة يعني أن عدداً كبيراً من الطالبات استطعن الإجابة عنها بنجاح. ولا يُعدّ ذلك عيباً في حد ذاته، لأن الأسئلة السهلة قد تكون ضرورية للتحقق من اكتساب المعارف الأساسية وتحقيق الأهداف التعليمية الأولية. غير أن ارتفاع نسبتها مع محدودية البنود الصعبة قد يقلل من قدرة الاختبار على إظهار الفروق بين الطالبات اللواتي يمتلكن مستويات تحصيل مرتفعة جداً واللواتي يحققن الحد الأدنى من الإتقان. لذلك، ينبغي ألا يستند الحكم على جودة توزيع الصعوبة إلى النسب وحدها، بل إلى مدى توافقها مع أهداف المقرر ومستوى الطالبات وطبيعة الاختبار النهائي.

وتتفق هذه النتيجة جزئياً مع ما توصلت إليه دراسة Nurlaeli (2024) ودراسة(2024) Roudhana ، اللتان أشارتا إلى غلبة البنود متوسطة الصعوبة في اختبارات اللغة العربية التي تناولتاها. ومع ذلك، فإن المقارنة تكشف أن تشابه وجود البنود المتوسطة لا يعني بالضرورة تطابق جودة الاختبارات؛ إذ تتأثر الجودة أيضاً بنسبة البنود السهلة والصعبة، وبقوة التمييز، وبمدى فعالية البدائل الخاطئة. وتتمثل الدلالة التربوية لهذه النتيجة في ضرورة بناء اختبار متوازن يضم بنوداً تقيس إتقان المعارف الأساسية، وأخرى تقيس الفهم والتطبيق، وبنوداً تتطلب عمليات معرفية أكثر تقدماً، وفقاً للأهداف التعليمية الفعلية، لا لمجرد تحقيق توزيع عددي متساوٍ بين مستويات الصعوبة.

أما من حيث قوة التمييز، فقد أظهرت النتائج أن ثلاثة بنود (12%) جاءت في فئة ممتازة، واثني عشر بنداً (48%) في فئة جيدة، وتسعة بنود (36%) في فئة مقبولة، وبنداً واحداً (4%) في فئة ضعيفة. ولم يُسجَّل أي بند ذي قوة تمييز سلبية. وبذلك، بلغ مجموع البنود المصنفة ضمن الفئتين الممتازة والجيدة 60% من إجمالي البنود، في حين بلغت نسبة البنود المقبولة والضعيفة 40%. وتكشف هذه النتائج أن غالبية البنود تمتلك قدرة جيدة أو مقبولة على التمييز بين الطالبات وفق مستويات أدائهن، إلا أن وجود 40% من البنود التي لا تتجاوز الفئة المقبولة، بما فيها بند ضعيف، يشير إلى مجال واضح للتحسين.

وتكتسب هذه النتيجة أهمية عند تفسير العلاقة بين مستوى الصعوبة وقوة التمييز؛ فالبند السهل جداً قد لا يميز بصورة كافية بين الطالبات إذا أجابت عنه معظمهن إجابة صحيحة، كما أن البند الصعب جداً قد يضعف تمييزه إذا أخفقت غالبية الطالبات في الإجابة عنه. ومع ذلك، لا توجد علاقة حتمية تجعل كل بند سهل ضعيف التمييز أو كل بند متوسط الصعوبة جيد التمييز؛ إذ تتأثر قوة التمييز أيضاً بوضوح صياغة السؤال، ومدى ارتباطه بالمحتوى المستهدف، وجودة البدائل، ومدى ملاءمة البند لمستويات الطالبات. ولذلك، ينبغي مراجعة البنود المقبولة والضعيفة في ضوء نتائج تحليلها التفصيلية، مع إعطاء الأولوية للبند ذي التمييز الضعيف، والتحقق من مفتاح الإجابة وصياغة السؤال ومدى توافقه مع المادة التي دُرّست.

أما عدم ظهور أي بند ذي تمييز سلبي، فيُعدّ مؤشراً إيجابياً؛ لأنه يعني عدم وجود بند أظهر النمط المعاكس المتوقع في التمييز بين مجموعات الأداء وفق معيار التحليل المستخدم. ومع ذلك، لا يكفي هذا المؤشر وحده للحكم على خلو الاختبار من المشكلات، بل ينبغي تفسيره إلى جانب بقية النتائج. وبناءً على ذلك، فإن تحسين جودة الاختبار يتطلب المحافظة على البنود الممتازة والجيدة بعد التحقق من توافقها مع الأهداف، ومراجعة البنود المقبولة، وإعادة بناء البنود الضعيفة أو استبعادها عند ثبوت عدم صلاحيتها.

فعالية المشتتات والتكامل بين المؤشرات: الدلالات التربوية لتحسين الاختبار

أظهر تحليل فعالية المشتتات أن 8% من البنود جاءت في فئة ممتازة، و36% في فئة جيدة، و36% في فئة متوسطة، و20% في فئة ضعيفة، في حين لم يُصنّف أي بند ضمن الفئة الرديئة. وتعني هذه النتائج أن المشتتات في 44% من البنود تقع ضمن الفئتين الممتازة والجيدة، بينما تقع نسبة 36% في الفئة المتوسطة، وتحتاج نسبة 20% إلى المراجعة بسبب ضعف قدرتها على جذب الطالبات. وتشير هذه الصورة إلى أن المشتتات تؤدي وظيفتها بدرجات متفاوتة، وأن جودة البدائل الخاطئة تمثل أحد الجوانب التي تحتاج إلى مزيد من العناية في تطوير الاختبار.

وتتمثل وظيفة المشتتات في أسئلة الاختيار من متعدد في توفير بدائل خاطئة تبدو معقولة للطالبات اللاتي لم يتقنّ المحتوى المستهدف، بحيث تساعد استجابتهن على الكشف عن التصورات غير الصحيحة أو أوجه القصور في الفهم. وعندما يكون المشتت واضح الخطأ أو بعيداً عن موضوع السؤال، فقد لا يجذب أي طالبة، مما يقلل من فاعلية البند في التمييز بين مستويات الأداء. ومن ثم، فإن ضعف المشتتات لا يرتبط بمجرد كونها إجابات خاطئة، بل بمدى معقوليتها واتصالها بالمحتوى وبالأخطاء التي قد تقع فيها الطالبات أثناء التعلم.

وفي ضوء ذلك، فإن نسبة المشتتات الضعيفة البالغة 20% تستدعي مراجعة البدائل المرتبطة بهذه البنود، من خلال التحقق من وضوح صياغتها، ومعقوليتها، وارتباطها بالمحتوى التعليمي، ومدى قدرتها على تمثيل الأخطاء الشائعة لدى الطالبات. ويمكن الاستفادة من نتائج تحليل الاستجابات لتحديد البدائل التي لم تخترها أي طالبة أو اختارتها نسبة محدودة منهن، ثم إعادة صياغتها بما يجعلها أكثر معقولية من الناحية التعليمية، دون الإخلال بوضوح الإجابة الصحيحة أو خلق أكثر من إجابة محتملة. ويُستحسن أيضاً الاستفادة من خبرة معلمات اللغة العربية في تحديد الأخطاء الشائعة لدى الطالبات عند بناء المشتتات الجديدة.

وعند النظر إلى المؤشرات الخمسة مجتمعةً، يتضح أن جودة الاختبار لا يمكن اختزالها في ارتفاع معامل الصدق أو الثبات وحدهما. فقد أظهر الاختبار معامل ارتباط بلغ 0,77 ومعامل ثبات بلغ 0,87، وهما

من الجوانب الإيجابية في النتائج، كما أن 60% من البنود جاءت في فئتي التمييز الممتاز والجيد، ولم تظهر بنود ذات تمييز سلبي. وفي المقابل، كشف التحليل عن عدم توازن نسبي في مستويات الصعوبة، إذ بلغت نسبة البنود السهلة 44% مقابل 8% فقط للبنود الصعبة، فضلاً عن وجود 20% من البنود ذات المشتتات الضعيفة، وبند واحد غير متوافق مع مؤشر جدول المواصفات. وتوضح هذه النتائج أن الاختبار يمتلك أساساً قياسياً جيداً، لكنه لا يزال يحتاج إلى تحسين بعض خصائص بنوده حتى تصبح جودة التقويم أكثر تكاملاً.

وتنسجم هذه القراءة مع المنظور الشامل للنظرية الكلاسيكية للاختبارات، الذي يقتضي دراسة خصائص البنود والاختبار معاً بدلاً من إصدار حكم نهائي استناداً إلى معامل منفرد. كما تتفق النتائج مع دراسة Febriyanti (2024) في إبراز أهمية فحص الصدق والثبات عند تقييم أسئلة اللغة العربية، وإن كانت الدراسة الحالية قد أظهرت قيماً أعلى في هذين المؤشرين وفق النتائج المتاحة. ومع ذلك، ينبغي توخي الحذر عند إجراء المقارنة المباشرة، لأن اختلاف العينات، وطبيعة الاختبارات، وإجراءات حساب المؤشرات، ومعايير التصنيف قد يؤثر في النتائج. وتضيف الدراسة الحالية إلى هذا النقاش عرضاً متكاملاً للمؤشرات الخمسة ضمن سياق تعليمي محدد، بما يتيح تحديد جوانب القوة والقصور في الاختبار بصورة أكثر شمولاً.

ويمكن، بناءً على هذه النتائج، ترتيب إجراءات التحسين بحسب الأولوية؛ إذ تبدأ بمراجعة البند غير المتوافق مع جدول المواصفات، ثم فحص البند ضعيف التمييز والبنود ذات المشتتات الضعيفة، وبعد ذلك مراجعة توزيع مستويات الصعوبة في ضوء أهداف المقرر ومحتواه. كما يمكن الاحتفاظ بالبنود الجيدة بعد التحقق من مناسبتها، وتوثيق نتائج تحليلها في بنك أسئلة يُحدّث بصورة دورية. ومن شأن اعتماد هذه الإجراءات أن يساعد معلمات اللغة العربية على الانتقال من إعداد الاختبارات بالاعتماد على الخبرة وحدها إلى إعدادها وتقويمها استناداً إلى أدلة كمية، مع مراعاة الأهداف التعليمية وخصائص المتعلمات.

الجدول 1. ملخص نتائج تحليل جودة بنود الاختبار النهائي

المؤشر	النتيجة	التفسير
الصدق (معامل الارتباط)	0,77	جيد وفق معيار التصنيف المستخدم
صدق المحتوى	بنداً صحيحاً (96%)، وبند 24 (4%) واحد غير صحيح	توافق مرتفع مع جدول المواصفات، مع ضرورة مراجعة بند واحد
الثبات	0,87	عالٍ جداً وفق معيار التصنيف المستخدم
مستوى الصعوبة	صعب، و48% متوسط، 8% و44% سهل	ميل نسبي نحو البنود السهلة، مع محدودية البنود الصعبة
قوة التمييز	ممتاز، و48% جيد، و36% 12% مقبول، و4% ضعيف	من البنود ممتازة أو جيدة، و40% 60% مقبولة أو ضعيفة
فعالية المشتتات	ممتاز، و36% جيد، و36% 8% متوسط، و20% ضعيف	ممتازة أو جيدة، مع حاجة خُمس 44% البنود إلى تحسين المشتتات

الخاتمة

خلصت الدراسة إلى أن جودة بنود الاختبار النهائي للفصل الدراسي في مادة اللغة العربية للصف الثاني المتوسط بمدرسة لقمان الحكيم الإسلامية ديلي سردانج كانت جيدة بوجه عام، مع وجود جوانب محددة تحتاج إلى التطوير. فقد بلغ معامل الارتباط 0,77، وصُنّف ضمن الفئة الجيدة، في حين بلغ معامل الثبات 0,87، وهو ما يشير إلى اتساق مرتفع في القياس وفق المعيار المستخدم. كما تبيّن أن 96% من البنود متوافقة مع مؤشرات جدول المواصفات، وأن 60% من البنود تقع ضمن فئتي التمييز الممتاز والجيد، دون تسجيل

أي بند ذي تمييز سلبي. وفي المقابل، أظهر توزيع الصعوبة ميلاً نحو البنود السهلة، التي بلغت نسبتها 44%، مقابل 8% فقط للبنود الصعبة، كما احتوت 20% من البنود على مشتتات ضعيفة، فضلاً عن وجود بند واحد غير متوافق مع مؤشر جدول المواصفات. وتؤكد هذه النتائج أن ارتفاع الصدق والثبات لا يغني عن مراجعة الخصائص الأخرى للبنود، وأن الحكم على جودة الاختبار ينبغي أن يستند إلى قراءة متكاملة لجميع المؤشرات.

وفي ضوء ذلك، توصي الدراسة معلمات اللغة العربية ومعدّي الاختبارات بمراجعة البند غير المتوافق مع جدول المواصفات، وإعادة فحص البنود ضعيفة التمييز، وتحسين المشتتات التي لا تجذب الطالبات بدرجة كافية. كما توصي بمراجعة توزيع مستويات الصعوبة بما يتناسب مع الأهداف التعليمية ومستويات التحصيل المستهدفة، دون فرض توزيع متساوٍ بصورة آلية. وعلى المستوى المؤسسي، يُستحسن أن تدعم المدرسة التحليل الدوري لبنود الاختبارات باستخدام برنامجANATES ، وأن تنشئ بنكاً رقمياً للأسئلة يتضمن نتائج تحليل البنود وتوصيات تطويرها، إلى جانب تعزيز تدريب المعلمات على إعداد الأسئلة وتحليل خصائصها الإحصائية. أما الدراسات المستقبلية، فيمكن أن توسع نطاق التحليل ليشمل اختبارات في صفوف ومواد تعليمية مختلفة، وأن تقارن نتائج التحليل الإحصائي بمراجعة الخبراء للمحتوى؛ بما يسهم في تعزيز جودة التقويم وتحسين ممارسات تعليم اللغة العربية.

المراجع

Aini, U., Lathifa, N., & Jelita, M. (2025). Anates: Analysis of Objective Test Item Quality. *Journal of Education and Islamic Contemporary Issue*, *1*(1), 45–59.

Arifin, Z. (2017). *Evaluasi Pembelajaran*. Remaja Rosdakarya.

Asyim, M. M., Ridlo, U., & Raswan. (2024). Validitas Tes Dalam Pembelajaran Bahasa Arab. *Journal of Arabic Studies*, *5*(4), 689–701. https://doi.org/10.37274/ukazh.v5i4.1208

Dianova, F. R., & Anwar, N. (2024). Analisis Butir Uji Validitas, Reliabilitas, Tingkat Kesukaran, Dan Daya Pembeda Soal Sumatif Bahasa Arab SD Islam. *Jurnal Bahasa Daerah Indonesia*, *1*(3), 13. https://doi.org/10.47134/jbdi.v1i3.2863

Farida, & Musyarofah, A. (2021). Validitas Dan Reliabilitas Dalam Analisis Butir Soal. *Al-Mu'arrib: Jurnal Pendidikan Bahasa Arab*, *1*(1), 34–44.

Febriyanti, S. A. S. (2024). *تحليل أسئلة الاختبار النهائي في مادة اللغة العربية لطلاب الصف الأول المتوسط بالمدرسة الجمعيّة الوصليّة 40 تينوكّة بمنطقة سردانج بداغي سومطرة الشماليّة.*

Handriawan, D., & Nurman, M. (2021). *Evaluasi Pembelajaran Bahasa Arab*. Sanabil.

Hermawan, A. (2021). *Penilaian Pembelajaran Bahasa Arab: Prinsip Dan Operasional*. Remaja Rosdakarya.

Mushofa, M., Hermina, D., & Huda, N. (2024). Memahami Populasi Dan Sampel: Pilar Utama Dalam Penelitian Kuantitatif. *Jurnal Syntax Admiration*, *5*(12), 5937–5948. https://doi.org/10.46799/jsa.v5i12.1992

Nasyrah, U., Bustanur, & Zulhaini. (2024). Analisis Butir Soal Pilihan Ganda Menggunakan Teori Pengukuran Klasik (Pada Penilaian Akhir Tahun (PAT) Mata Pelajaran PAI Dan BP Kelas VIII Di SMPN 2 Kuantan Mudik). *Jurnal Al-Hikmah*, *6*(2), 56–66.

Nurlaeli, S. (2024). *Analisis Butir Soal Sumatif Tengah Semester (STS) Mata Pelajaran Bahasa Arab Siswa Kelas VIII MTs Ma'arif NU 1 Patikraja Banyumas*.

Roudhana. (2024). *تحليل بنود أسئلة اللغة العربية في المدرسة المتوسطة بمعهد دار الحكمة*.
Savika, H. I., & Zuhriyah, I. A. (2024). Peran Analisis Butir Soal Terhadap Kualitas Soal, Kompetensi Guru, Dan Prestasi Belajar Peserta Didik Di Sekolah Dasar. *PANDU: Jurnal Pendidikan Anak Dan Pendidikan Umum*, *2*(2), 43–51. https://doi.org/10.59966/pandu.v2i2.856
Sihombing, P. R., Arsani, A. M., Lameky, V. Y., Hapsery, A., & Astuti, M. (2024). *Analisis Butir Soal Dalam Berbagai Software*. Minhaj Pustaka.
Sudaryono. (2019). *Metode Penelitian Kuantitatif, Kualitatif, Dan Mix Method*. Rajagrafindo Persada.
Sugiyono. (2023). *Metode Penelitian Pendidikan (Kuantitatif, Kualitatif, Kombinasi, R&D, Dan Penelitian Tindakan)*. Alfabeta.
Sumaryanta. (2021). *Teori Tes Klasik Dan Teori Respon Butir: Konsep Dan Contoh Penerapannya*. Confident.
Warju, Ariyanto, S. R., Soeryanto, & Trisna, R. A. (2020). Analisis Kualitas Butir Soal Tipe Hots Pada Kompetensi Sistem Rem Siswa Di Sekolah Menengah Kejuruan. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *17*(1), 95–104.
Wiguna, S. (2021). *Aplikasi ANATES Dalam Evaluasi Pembelajaran*. Pena Persada.
بيهقي, م., المبارك, أ. ج., & ربانية, ن. (2025). *التقويم في تدريس اللغة العربية*. مطبعة جدس.