

تحليل بنود الأسئلة في اختبار المفردات التحريري

بمعهد دار اللغة والدعوة للبنات

Ardianti, M.Pd.I

Universitas Islam Internasional Darullughah wadda'wah

dian11ardianti@gmail.com

الملخص

مشكلة هذا البحث هي ما درجة صعوبة الأسئلة وتميزها ومتابعتها في اختبار المفردات التحريري للفصل الثاني العالي المعهدي "ب" بمعهد دار اللغة والدعوة للبنات للعام الدراسي 2025 م. ويهدف هذا البحث لمعرفة جودة بنود الأسئلة من حيث معامل الصعوبة والتميز ومتابعتها. استخدم هذا البحث المنهج الكمي الوصفي معتمداً على أسلوب تحليل بنود الأسئلة. وقد جمعت البيانات من نتائج اختبار المفردات الذي خضع له 78 طالبة من الفصل الثاني العالي المعهدي "ب". وتم تحليل كل البنود باستخدام Anates V.4 لتحديد معامل الصعوبة والتميز لكل سؤال. أظهرت نتائج التحليل أن من بين 27 سؤالاً تم تحليله، كانت هناك أسئلة صعبة تصل إلى 3 أسئلة (11%)، ومتوسطة تصل إلى 17 سؤالاً (63%)، وسهلة تصل 7 أسئلة (26%) حسب معامل الصعوبة. ومن حيث قوة التمييز، كانت هناك أسئلة قبيحة تصل إلى سؤال واحد (3،7%)، وكافية تصل إلى 5 أسئلة (5،18%)، وجيدة تصل إلى 17 سؤالاً (63%)، وجيدة جداً تصل إلى 4 (8،14%). وهذه النتائج تدل على أن بعض الأسئلة تحتاج إلى التنقيح لتكون أكثر فاعلية في قياس قدرات الطلاب بشكل عادل. ويؤمل أن يسهم هذا البحث في تحسين جودة تقييم تعليم اللغة العربية، وخاصة في إعداد أسئلة المفردات التحريرية التي تتميز بالصدق والموثوقية.

الكلمات المفتاحية : اختبار المفردات؛ بنود الأسئلة؛ تحليل.

Abstract

This study aims to evaluate the quality of vocabulary test items administered to second-year students of Aliyah Ma'had "B" at Ma'had Darul Lughah wa ad-Da'wah lil Banat in the 2025 academic year. The research focuses on three main aspects: difficulty level, discriminating power, and overall item performance. Using a descriptive quantitative method, data were collected from a vocabulary test taken by 78 students, and all 27 items

were analyzed with Anates V.4 to determine their psychometric characteristics. The results show that, based on the difficulty index, 11% of the items were categorized as difficult, 63% as moderate, and 26% as easy. Regarding discriminating power, 3.7% of the items were considered poor, 18.5% fair, 63% good, and 14.8% very good. These findings indicate that most items function effectively, although several require revision to better align with assessment standards. Items with weak performance should be improved or rewritten, while items showing strong psychometric quality can be stored in the test bank for future use. This study is expected to contribute to enhancing the quality of Arabic language assessment, particularly in developing valid and reliable vocabulary test items.

Keywords: Analysis; Mufrodat Exam; Test Items.

المقدمة

ولمعرفة ما إذا كانت الأهداف التعليمية قد تحققت، فلا بدّ من إجراء الاختبارات المناسبة. يجرى تحليل بنود الأسئلة لتوحيد الأسئلة التي تم إعدادها، وذلك للحصول على بنود ذات جودة عالية. ولا بدّ من تحليل كل بند من بنود الاختبار قبل اعتباره أداة قياس جيدة. إذ تُشير نتائج البحوث إلى أهمية فحص مستوى الصعوبة، وقوة التمييز، وفعالية البدائل المضللة لضمان جودة الاختبار (مولوجي، 2020).

كما هو معلوم لدينا، إنّ تحليل بنود الأسئلة يعدّ أحد أجزاء النقاش في مجال تقييم التعليم. ولذلك، قبل ان نناقشه بشكل أوسع، من الحكمة تذكير بمعنى التقييم. تشير كلمة "التقييم" في القاموس الكبير للغة الإندونيسية إلى الحكم أو التقدير. أما التقييم نفسه، حسب تعريف (نغالم بوروانتو، 2012) (2019) التقييم هو عملية التخطيط والحصول على توفير

في مجال التعليم، يعدّ التقييم جزءًا من عملية التعليم والتعلم، إذ يوضح دوره المهم في دعم تحقيق الأهداف التعليمية وتعزيز جودة التدريس (أمامي، 2018). فإن نظام التقييم الجيد يساهم في تشجيع المعلمين على استخدام استراتيجيات تدريس أفضل، كما يحفظ الطلاب على التعلم بجدّ واجتهاد.

وعادة ما يبدأ التقييم بعملية القياس التي تُعدّ الأساس في جمع البيانات حول تعلم الطلاب، إذ تُعدّ عملية القياس خطوة أولى في أي عملية تقييم تربوي (بوروانتو، 2016). والقياس فرع من فروع الإحصاء التطبيقي الذي يهدف إلى وضع الأسس لتطوير اختبارات أفضل بحيث تنتج اختبارات تعمل على نحو مثالي، وتكون صادقة وموثوقة (ساراغيه، 2019).

من 27 سؤالاً مقالياً. ويبلغ عدد الطلاب في الفصل الثاني العالي المعهدي "ب" للعام الدراسي 2025 م 78 طالبة.

المعلومات الضرورية للغاية-لاتخاذ قرارات بديلة قرار لنجاح الطلاب في برنامج معين. على سبيل المثال، النجاح في مادة معينة.

ومن الحقائق التي تم الحصول عليها من نتائج المقابلة مع مشرفة قسم البحوث التابع من قسم اللغة العربية في معهد دار اللغة والدعوة للبنات، وجدت الباحثة مشكلات في إجراء اختبار المفردات الذي يقام تنفيذه في كل الفصل الدراسي الواحد. منها استخدام المفردات الصعبة ونادرة الاستعمال في إعداد الأسئلة حتى يصعب حفظها للطلابات تماما (تاريغان، 2011؛ رحماواني، 2019)، والطلابات الراسبات في الاختبار أكثر من الناجحات لعدم بلوغهن معيار نجاح الاختبار، يعني ٧٠، وعدم التحليل حول أسئلة الاختبار لعدم قدرتهن عليه (مولوجي، 2020؛ إلفيانا، 2020). وتشير بعض الدراسات أيضا إلى أهمية تحليل مستوى الصعوبة وملاءمة المفردات مع قدرات المتعلمين حتى لا يؤدي ذلك إلى انخفاض نتائج الاختبار (سافيرا وأوتامي، 2021؛ فاوجري، 2020).

ويجب في تحليل البنود مراعاة مدى استيعاب الطلاب، ومستوى صعوبة الأسئلة، وقوة التمييز، ووظيفة البدائل المضللة، وكذلك الصدق والموثوقية. ويهدف ذلك إلى ضمان أن تكون الأسئلة متناسبة مع مستوى استيعاب الطلاب، وأن تكون مناسبة من حيث درجة الصعوبة، وأن تكون صادقة. ومن ثمّ يمكن تحقيق الأهداف التعليمية المنشودة. وتشير نتائج بعض الدراسات إلى أهمية تحليل مستوى الصعوبة، وقوة التمييز، وفعالية البدائل المضللة لضمان جودة البنود قبل اعتمادها في الاختبار (فاطمة والأفتح، 2019). ويُستخدم هذا التحليل لتقييم الاختبارات التي أُعدّت سواء من قبل المعلّم أو من خلال الاختبارات المعيارية الصادرة التي تضعها لجنة متخصصة، وذلك لضمان صلاحية البنود ودقّتها في قياس ما ينبغي قياسه (عبد الله، 2025)

وقد تمّ تطبيق هذا الاختبار في اختبار المفردات التحريري للفصل الثاني العالي المعهدي "ب" بمعهد دار اللغة والدعوة للبنات للعام الدراسي 2025 م ويتكون هذا الاختبار نظرا على هذه الخلفية، تهتم الباحثة بدراسة التقويم من خلال إجراء البحث حول تحليل بنود الأسئلة من حيث معامل الصعوبة

بعض الدراسات المتعلقة بذلك (مولوجي، 2020؛ سابتر وآخرون، 2020). لأن نتيجة الاختبار غالباً تستخدم كمؤثر رئيسي في تقييم نجاح التدريس (سودجانا، 2017؛ بوروانتو، 2016). لذلك، من المهم تحليل على العاملين الرئيسين في اختبار المفردات، يعني معامل الصعوبة ومعامل التمييز لأجل خلق تقويم أكثر عدلاً ودقة في تقييم قدرات الطلاب (أزوار، 2019؛ أريكونتو، 2012). وبهدف المساهمة في تطوير جودة التعليم والتقييم وإعطاء صورة عملية واضحة عن مدى تحقيق اختبارات اللغة العربية للأهداف المنشودة في العملية التربوية.

منهجية البحث

إن النوع هذا البحث هو النوع التقويمي، أي البحث الذي يهدف إلى تقييم تنفيذ البرنامج (أريكونتو وجابار، 2009؛ ويران، 2011). إجراء التقويم في هذا البحث على بنود أسئلة اختبار المفردات من حيث معامل الصعوبة والتمييز في الفصل الثاني العالي المعهدي "ب" بمعهد دار اللغة والدعوة للعام الدراسي 2025 م لتحديد جودة الأسئلة بمدخل كمي. لذلك تتم الاستعانة بنتائج التحليل الكمي لبنود الاختبار كما أشارت

كما في KBBI أن أفراد البحث هي الأشخاص أو الأشياء التي يتم ملاحظتها في سياق استهدافها. الأفراد في هذا البحث كانت جميع الطالبات الجالسات في الفصل الثاني العالي المعهدي "ب" بمعهد دار اللغة والدعوة للعام الدراسي 2024-2025 م، يعني 78 طالبة. وتستعمل المتغير التقدي (*control variable*) هو متغير يتم التحكم فيه أو تحويله إلى ثبات بحيث لا يتأثر تأثير المتغير المستقل على المتغير التابع بعوامل الخارجية التي لا يتم فحصها سوغيونو، (2013).

تم الحصول على بيانات البحث باستخدام طريقة التوثيق التي تشمل ورقة الأسئلة في اختبار المفردات التحريري للفصل الثاني العالي المعهدي "ب" بمعهد دار اللغة والدعوة للبنات للعام الدراسي 2025 م، ومفتاح الأجوبة، وأوراق عمل الطالبات. كما أشار بعض الباحثين إلى أن طريقة التوثيق تُعد من

الأساليب المناسبة لجمع البيانات في البحث التربوي (أريكونتو، 2013؛ سوغيونو، 2016)

وبعد ذلك تم تحليل البيانات باستخدام برنامج **Anates V.4** لتحديد جودة بنود الأسئلة من حيث معامل الصعوبة والتمييز، وهي عملية تم توضيحها في عدة دراسات استخدمت البرنامج نفسه في تحليل جودة الاختبار (إلفيانا، 2020؛ كوسوماواتي، 2017؛ مولوجي، 2020؛ سابترا وآخرون، 2022).

لتحليل البيانات يستخدم تفسير نتائج حسابات معامل الصعوبة المعيار التالي:

معامل الصعوبة	صفات	متابعة
- 0.00 0.30	صعبة	مرفوضة
- 0.31 0.70	متوسطة	مقبولة
- 0.71 1.00	سهلة	مرفوضة

وكان تفسير نتائج حسابات معامل التمييز المعيار التالي:

معامل التمييز	صفات	متابعة
- 0,00 0,20	قبيحة	مرفوضة
- 0,21 0,40	كافية	مرفوضة
- 0,41 0,70	جيدة	مقبولة
- 0,71 1,00	جيدة جدا	مقبولة
سلي	كل شيء ليس جيدا	مرفوضة

المتابعة التي يمكن إجراؤها للمعلمين بعد معرفة جودة الأسئلة كما يلي :

- بنود الأسئلة ذات الجودة الجيدة تستطيع أن يدخلها في بنك الأسئلة، حتى يمكن استخدامها للاختبار التالي.
- بنود الأسئلة ذات الجودة الناقصة لم تستطع أن يدخلها في بنك الأسئلة بل يحتاج إلى المراجعة والتنقيح مرة أخرى.
- بنود الأسئلة ذات الجودة السيئة لا تستطيع أن يدخلها في بنك الأسئلة ولا بد يدها بسؤال جديد.

نتائج البحث ومناقشتها

1. جودة بنود الأسئلة في اختبار المفردات

التحريري من حيث معامل الصعوبة

معامل الصعوبة هو فرصة إجابة السؤال بشكل صحيح على مستوى القدرة المعينة التي تذكر بشكل مؤشر. يتم حساب مؤشر مستوى الصعوبة لكل رقم سؤال (أريكونتو، 2012). يتم رؤية مستوى الصعوبة غرض جودة Anates V.4 من خلال قيمة الغرض. جودة الأسئلة الجيدة لها المشكلات. ومن ناحية أخرى، الأسئلة الصعبة للغاية ستؤدي إلى عدم حماسة الطلاب لحل الأسئلة، لأنها بعيدة عن متناول قدرات الطلاب (أريكونتو، 2012). عادة ترتبط وظيفة معامل صعوبة البنود بأهداف الاختبار، لأن مستوى الصعوبة ينبغي أن يعكس طبيعة الاختبار والغرض منه (أريكونتو، 2012). الأسئلة المستخدمة في الامتحانات الفصلية تكون درجة صعوبتها متوسطة. يتراوح مؤشر مستوى الصعوبة للأسئلة الجيدة بين 0.31 – 0.70 (أريكونتو، 2012).

في تطبيق التحليل الإحصائي، تُظهر نتائج معامل صعوبة الأسئلة في اختبار المفردات التحريري للفصل الثاني العالي المعهدي "ب"

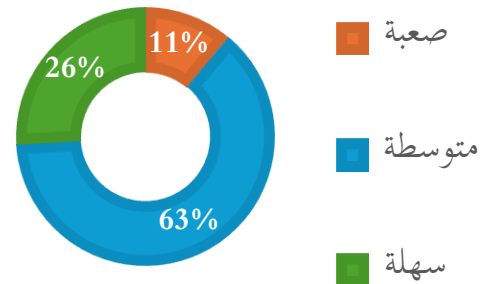
للعام الدراسي 2025 م بمساعدة برنامج Anates V.4 أن الأسئلة المصنفة على أنها صعبة تصل إلى 3 أسئلة (11%) وهي 15، 18، 20، والأسئلة السهلة تصل إلى 7 أسئلة (26%) وهي 3، 4، 10، 12، 14، 21، 24. ويمكن إسقاط هذه الأسئلة أو إعادة فحصها لأسباب متعددة، مثل الصياغة غير الواضحة أو عدم التركيز على الهدف أو استعمال مفردات نادرة وغير مناسبة للتقييم. وبعد ذلك، يمكن تحسين البنود بحيث تُستخدم مرة أخرى في الاختبار التالي.

أما الأسئلة المتوسطة فتصل إلى 17 سؤالاً (63%) وهي 1، 2، 5، 6، 7، 8، 9، 11، 13، 16، 17، 19، 23، 24، 25، 26، 27، ويمكن إدراجها في بنك الأسئلة لأنها واضحة وتستطيع قياس نتائج التعلم بشكل فعال (أزوار، 2019).

معامل الصعوبة	عدد الأسئلة	نسبة المئوية	رقم السؤال
صعبة	3	11	15، 18، 20

سهلة	7	26	4، 3، 10، 12، 14، 21، 24
مت وسطة	1 7	6 3	1، 5، 2، 7، 6، 9، 8، 11، 13، 16، 17، 19، 22، 23، 25، 26، 27

توزيع الرسم البياني الدائري لبنود
الأسئلة من حيث معامل الصعوبة



تُظهر نتائج تحليل بنود اختبار المفردات أنّ معظم الأسئلة تقع في مستوى الصعوبة المتوسطة، مما يدلّ على ملائمتها لقياس قدرات الطالبات وانسجامها مع أهداف التعلم (أريكونتو، 2012؛ سودجانا، 2017). كما تظهر بعض البنود التي تحتاج إلى تحسين بسبب سهولتها أو صعوبتها الزائدة. وبشكل عام، يتّضح أنّ جودة الاختبار جيدة، مع الحاجة إلى تعديلات محدودة لضمان دقة القياس وعدالته (أزوار، 2019).

وبناءً على ذلك، توصي الباحثة بمراجعة البنود السهلة والصعبة لمعرفة أسباب الخلل وتعديل صياغتها أو بدائلها،

ثم إعادة اختبارها بعد التعديل (مولوجي، 2020). كما توصي بالاعتماد على البنود الجيدة في بنك الأسئلة واستخدام التحليل الإحصائي لبنود الاختبار بشكل منتظم لضمان جودة مستقرة للتقويم (سابترا وآخرون، 2022؛ أريكونتو، 2012).

2. جودة بنود الأسئلة في اختبار المفردات التحريري من حيث معامل التمييز

معامل التمييز هي قدرة السؤال على التمييز بين الطلاب الذين أتقنوا المادة والطلاب الذين لم يتقنوا المادة التي يتم اختبارها (كوسيري، 2012). معرفة معامل التمييز لبنود الأسئلة من أمر مهم، لأن أحد الأساسيات في إعداد بنود الاختبار لنتائج التعلم هو الافتراض بأن قدرات أحد المختبرين تختلف عن الآخر. يجب أن تكون بنود اختبار نتائج التعلم قادرة على تقديم نتائج الاختبار التي توضح الاختلافات في القدرات بين الطلاب (أزوار، 2019). الأسئلة الجيدة هي البنود التي تمكن أن تميز الطلاب الأذكياء عن الطلاب الأقل ذكاء. وفي هذه الحالة، يمكن للطلاب الأذكياء الإجابة على الأسئلة بشكل صحيح (أريكونتو، 2012).

وفقًا لما ورد في أدبيات تحليل بنود الاختبار لدى سوهرسيمي أريكونتو، تظهر نتائج التحليل لتلك الأسئلة بمساعدة برنامج Anates V.4 أنها قبيحة تصل إلى سؤال واحد (3،7%) وهي 4، وكافية تصل إلى 5 أسئلة (18،5%) وهي 3، 5، 8، 18، 20، فيمكن تتبعها ثم تحسينها مرة أخرى في اختبار نتائج التعلم في المستقبل لمعرفة إذا كان معامل التمييز يزداد أم لا أو يتم التخلص منها (أريكونتو، 2012).

أما الأسئلة الجيدة فتصل إلى 17 سؤالاً (63%) وهي 1، 2، 6، 7، 9، 10، 11، 12، 13، 14، 15، 16، 17، 19، 21، 22، 24، والجيدة جدًا تصل إلى 4 (14،8%) وهي 23، 25، 26، 27، فيمكن تخزين البنود بقدرة تمييز جيدة في بنك الأسئلة، ويمكن إصدار هذه البنود مرة أخرى خلال اختبار نتائج التعلم التالي (أريكونتو، 2013؛ إلفيانا، 2020).

تشير نتائج تحليل بنود اختبار المفردات إلى أن معظم الأسئلة تمتلك مستوى صعوبة مناسبًا وقدرة تمييز جيدة، مما يدل على أن الاختبار قادر بشكل عام على قياس قدرات الطالبات بصورة عادلة وفعالة. ومع ذلك،

الثاني العالمي المعهدي "ب" للعام الدراسي 2024 م من حيث معامل الصعوبة والتميز أن هناك 19 سؤالاً بجودة جيدة (70,4%)، و 8 أسئلة بجودة ناقصة (29,6%)، وبجودة سيئة 0 (0%). (2019).

توصي الباحثة بما يلي:

1. مراجعة البنود الضعيفة من حيث الصعوبة أو التمييز، وإعادة صياغتها أو تحسين بدائلها بحيث تتوافق مع أهداف التعلم (أريكونتو، 2012).

2. اعتماد البنود الجيدة في بنك الأسئلة لاستخدامها في الاختبارات المستقبلية، لأنها قادرة على قياس التعلم بكفاءة (مولوجي، 2020).

3. إجراء تحليل مستمر لبنود الاختبارات بعد كل فصل دراسي لضمان جودة التقييم وتطوير أدوات القياس بشكل مستمر (إلفيانا، 2020).

3. متابعة جودة بنود الأسئلة في اختبار المفردات التحريري من حيث معامل الصعوبة والتمييز

ظهرت نتائج تحليل بنود الأسئلة في اختبار المفردات التحريري للفصل

ممكن إجراء المتابعة الملموسة من قبل قسم البحوث باعتباره منشئ أسئلة الامتحان الفصلي. كالباحثة، المتابعة التي يمكنها القيام بها بعد معرفة جودة الأسئلة هي تقديم وإبلاغ قسم البحوث بنتائج البحث. إجراءات المتابعة الملموسة التي يمكن للمعلمين اتخاذها بعد معرفة جودة بنود الأسئلة هي كما يلي:

- بنود الأسئلة ذات الجودة الجيدة (70,4%) وهي 2، 5، 6، 7، 9، 11، 13، 15، 16، 17، 19، 21، 22، 23، 24، 25، 26، 27، 28، فتستطيع أن تُدرج في بنك الأسئلة لأنها صالحة لقياس نتائج التعلم بشكل فعال (أريكونتو، 2012؛ عارفين، 2012).

- بنود الأسئلة ذات الجودة الناقصة (29,6%) وهي 3، 4، 8، 10، 12، 14، 18، 20، فلا يمكن

يصعب التمييز بين الطلاب ذوي القدرة المنخفضة أو العالية (سودجانا، 2017) بناء على معامل الصعوبة والتمييز، العلاقة بين الاثنين هي الأكثر هيمنة في معامل التمييز، لأنه إذا كان هناك سؤال شديد الصعوبة أو سهل للغاية، ولكن له معامل التمييز تفي بالمعايير، فيمكن اختيار ذلك السؤال وإقباله كبديل لتخزينها في بنك الأسئلة (إلفيانا، 2020). وإذا كانت هناك أسئلة شديدة الصعوبة وأسئلة سهلة للغاية وكان معامل التمييز لا تفي بالمعايير، فيجب مراجعتها وإعادة صياغتها (أريكونتو، 2012؛ أزوار، 2019).

الخلاصة

تُظهر نتائج تحليل اختبار المفردات أنّ جودة البنود يمكن تقييمها من ثلاثة جوانب أساسية. أولاً، من حيث معامل الصعوبة، يتضح أنّ أغلبية البنود تقع في المستوى المتوسط، مما يشير إلى ملاءمتها لقياس قدرات الطالبات بشكل متوازن، في حين تحتاج البنود السهلة والصعبة إلى مراجعة لضمان عدالة

إدراجها في بنك الأسئلة بل تحتاج إلى المراجعة والتنقيح لتحسين وضوحها وارتباطها بالأهداف التعليمية (أزوار، 2019؛ إلفيانا، 2020).

بنود الأسئلة ذات الجودة السيئة (0 فلا يمكن استخدامها مرة أخرى ويجب استبدالها بنود جديدة تتوافق مع معايير القياس الجيد (أريكونتو، 2012؛ عارفين، 2012).

كلما زاد مؤشر مستوى الصعوبة الذي تم الحصول عليه من نتائج الحساب، كان مستوى صعوبة السؤال أسهل. وكلما كان مستوى الصعوبة أصغر، زاد مستوى صعوبة السؤال أصعب (أريكونتو، 2012). كلما ارتفع مؤشر معامل التمييز للسؤال، يعني أن السؤال أكثر قدرة على تمييز الطلاب الذين فهموا المادة عن الطلاب الذين لم يفهموا المادة (أزوار، 2019).

يؤثر معامل صعوبة السؤال بشكل كبير على قدرة الطلاب على إجابة السؤال. الأسئلة التي تحتوي على مؤشر صعوبة كبير، أي مستوى الصعوبة السهل، ستسهل على الطلاب الإجابة على الأسئلة. هذه السهلة سيصعب اكتشاف معامل التمييز للأسئلة، لأنه

القياس. ثانيًا، من حيث معامل التمييز، تُظهر معظم البنود قدرة جيدة على التمييز بين الطالبات اللاتي يفهمن المادة واللاتي لا يفهمنها، مما يعكس جودة مقبولة لأداة التقييم، بينما تحتاج البنود ذات التمييز الضعيف إلى تعديل أو إعادة صياغة. ثالثًا، فيما يتعلق بمتابعة جودة البنود، تؤكد النتائج أهمية تحسين البنود الناقصة واعتماد البنود الجيدة في بنك الأسئلة، مع استمرار استخدام التحليل الإحصائي كأداة أساسية لتطوير جودة الاختبارات في المستقبل

قائمة المراجع

- Abdullah, I., Listyotami, M. K., Ferita, R. A., Lisa, L., Maulida, D., Yanti, W., & Hidayat, F. (2025). Analisis kualitas butir soal asesmen sumatif matematika pilihan ganda semester genap kelas VIII MTs Baitul Ma'mur tahun ajaran 2023/2024. *Nusantara Journal of Education and Social Science*, 2(1), 27–32. <https://doi.org/10.69959/nujess.v2n1.49>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan*. Jakarta: Bumi Aksara.
- Arikunto, S., & Jabar, C. S. A. (2009). *Evaluasi program pendidikan: Pedoman teoretis praktis bagi mahasiswa dan praktisi pendidikan*. Jakarta: Bumi Aksara.
- Azwar, S. (2019). *Penyusunan skala psikologi*. Yogyakarta: Pustaka Pelajar.
- Elviana, E. (2020). Analisis butir soal evaluasi pembelajaran PAI menggunakan ANATES. *Jurnal Mudarrisuna*, 10(2), 209–224.
- Fatimah, L. U., & AlFath, K. (2019). Analisis kesukaran soal, daya pembeda, dan fungsi distraktor. *Al-Manar: Jurnal Komunikasi dan Pendidikan Islam*, 8(2).
- Fajri, N. (2020). Faktor-faktor yang mempengaruhi hasil belajar siswa pada pembelajaran bahasa. *Jurnal Pendidikan dan Pembelajaran*, 7(1), 55–63.
- Kusumawati, D. (2017). Analisis kualitas soal ujian akhir sekolah menggunakan program ANATES. *Jurnal Pendidikan dan Pengajaran*, 50(1), 45–54.
- Muluki, A. (2020). Analisis kualitas butir tes semester ganjil mata pelajaran IPA kelas IV MI Radhiatul Adawiyah. *Jurnal Ilmiah Sekolah Dasar*, 4(1), 86–96. <https://doi.org/10.23887/jisd.v4i1.23335>
- Purwanto, N. (2016). *Prinsip-prinsip dan teknik evaluasi pengajaran*. Bandung: Remaja Rosdakarya.
- Rahmawati, I. (2019). Analisis kesulitan belajar siswa dalam penguasaan kosakata bahasa asing. *Jurnal Pendidikan Bahasa*, 5(2), 123–131.
- Rona, R. (2018). Pengukuran dan penilaian dalam evaluasi pembelajaran. *Cross-Border*, 1(1), 68–75.
- Saputra, H. D., Purwanto, W., Setiawan, D., Fernandez, D., & Putra, R. (2022). Hasil belajar mahasiswa: Analisis butir soal tes. *Edukasi: Jurnal Pendidikan*, 20(1), 15–27.

<https://doi.org/10.31571/edukasi.v20i1.3432>

- Saragih, A. A. (2019). *Pengantar konstruksi alat ukur*. Medan: Universitas Negeri Medan.
- Savira, D., & Utami, P. (2021). Analisis tingkat kesukaran dan daya pembeda pada soal pilihan ganda. *Jurnal Evaluasi Pendidikan*, 12(1), 45–55.
- Sumanto. (2014). *Statistik dalam penelitian pendidikan dan ilmu sosial*. Yogyakarta: CAPS.
- Sugiyono. (2016). *Metode penelitian kuantitatif, kualitatif dan R&D*. Bandung: Alfabeta.
- Sudjana, N. (2017). *Penilaian hasil proses belajar mengajar*. Bandung: Remaja Rosdakarya.
- Tarigan, H. G. (2011). *Pengajaran kosakata*. Bandung: Angkasa.
- Wirawan, W. (2011). *Evaluasi: Teori, model, standar, aplikasi dan profesi*. Jakarta: Rajawali Pers.
- Umami, M. (2018). Penilaian autentik pembelajaran Pendidikan Agama Islam dan Budi Pekerti dalam Kurikulum 2013. *Jurnal Kependidikan*, 6(2), 222–232. <https://doi.org/10.24090/jk.v6i2.2259>