DAS ENTWICKELN DES TESTINSTRUMENTS FÜR DEN DEUTSCHUNTERRICHT NACH DEM CURRICULUM 2013 ZUM THEMA "ALLTAGSLEBEN"

Ryan Nuansa Dirga
ryandirga@gmail.com
Primardiana Hermilia Wijayati
primardiana.hermilia.fs@um.ac.id
Iwa Sobara
iwa.sobara.fs@um.ac.id

ABSTRACT: This research is aimed to develope a test instrument based on curriculum 2013 with *Alltagsleben* theme for XIth Language class program in second semester. The development strategy in this research is strategy by Cennamo and Kalk. The developed test is a test to assess the reading comprehension, grammar and vocabulary ability of the students. There are some kinds of texts in the test that have different word amount. The short texts consist of 20 until 47 words and the students need 30 until 45 seconds to understand each text. Besides, the moderate texts consist of 57 until 82 words. The students need a minute to understand each text. The long texts consist of 114 until 126 words. The students need two minutes to understand each text. After the test has been developed, the test and items quality have been reviewed by experts and a trial has been held to check the readability, item difficulty, item discrimination, distractors' function and reliability of the test. The result is a test that has 54 items with good quality, 0,71 coefficient of validity and 0,863 coefficient of reliability.

Keywords : development, test instrument, curriculum 2013, German

Einführung

Das Curriculum spielt eine wichtige Rolle für die Bildung. Ein Curriculum ist ein Planungs- und Regelungssystem, das Lernziele, Inhalt, Unterrichtsmaterialien und auch die benutzten Lernmethoden umfasst. Das Curriculum wird als Grundlage des Bildungssystems benutzt, um die nationalen Bildungsziele zu erreichen (UU No.20/2003). Das Curriculum in dem indonesischen Bildungssystem verändert sich auch von Zeit zu Zeit. In dem indonesischen Bildungssystem galt des Curriculums beispielsweise vom Jahr 1994, 2004, 2006, sowie das noch neue geltende Curriculum 2013.

Das Curriculum 2013 ist eigentlich nicht wirklich neu. Das ist eine Entwicklung vom Curriculum 2006. Das vorherige Curriculum hat einige Nachteile, deshalb wird es verbessert. Die verbesserten Aspekte sind u.a. der Inhalt, die Kompetenzen, der Lernprozess, und die Bewertung (Mulyasa, 2013:60). Der Inhalt des Curriculums 2006 ist zu voll. Es wird durch die Zahl der Kurse gezeigt. Es gibt zu viele Kurse, deshalb werden sie im Curriculum 2013 gemischt. Und die meisten Kompetenzen im Curriculum 2006 sind auf die kognitiven Aspekte (Kenntnisse) konzentriert. Im Curriculum 2013 sind die

Grundkompetenzen nicht nur im Bereich der Kenntnisse, sondern auch des Verhaltens und der Fertigkeiten. Im Bildungsaspekt des Curriculums 2013 gibt es Charakter-Bildung. Die Leistung der SchülerInnen wird durch ihre Kompetenz bewertet. Auβer den oben genannten verbesserten Aspekten gibt es noch Unterschiede bei der Anwendung des Curriculums 2013. Bei der Anwendung des Curriculums 2006 mussten die LehrerInnen einen Syllabus entwerfen. Danach entwickelten sie den konzipierten Syllabus in einer Unterrichtsplanung. Im Curriculum 2013 ist es anders.

Bei der Anwendung des Curriculums 2013 brauchen die LehrerInnen den Syllabus nicht selbst zu entwerfen, weil er schon vorgegeben wird. Im vorgegebenen Syllabus (Permendikbud, 2013:2391) sind 4 Komptenzen vorhanden. Dies sind Hauptkompetenz 1 (*KI 1*), Hauptkompetenz 2 (*KI 2*), Hauptkompetenz 3 (*KI 3*), Hauptkompetenz 4 (*KI 4*). Diese vier Kompetenzen werden in Grundkompetenzen dargelegt. Die Grundkompetenzen werden in einem Lernprozess beigebracht.

Am Ende des Lernprozesses muss die Bewertung durchgeführt werden. Nurgiyantoro (2009:5) hat erwähnt, dass Bewertung ein Prozess ist, damit man es wissen kann, ob die Ziele bzw. die Lernziele schon erreicht sind oder nicht. Bei einem Interview mit den DeutschlehrerInnen haben sie zugegeben, dass sie Schwierigkeiten bei der Erstellung des Testes nach dem Curriculum 2013 haben. Der Grund ist, dass das Curriculum 2013 noch neu ist. Deswegen haben sie nicht genügende Information über das Curriculum 2013 erhalten. Die Situation beeinflüsst die Qualität des Testes.

Nach der Untersuchung von Lestari (2014) hat der von den LehrerInnen gebaute Deutschtest noch keine gute Qualität. Trotz der guten Zuverlässigkeit hat der Test niedrige Inhalt-Gültigkeit. Die Items von dem Test sind meistens einfach und mittelmäβig. Es gibt kein schwieriges Item. Die meisten Items sind im kognitiven Bereich C1 (Wissen). Die Distraktoren funktionieren auch nicht gut. Zusammenfassend hat der Test noch keine gute Qualität.

Wegen der schlechten Qualität des Testes kann der Test die Beherrschung der SchülerInnen nicht gut messen. Deshalb ist es sehr schwierig zu wissen, ob die SchülerInnen die Kompetenzen gut beherrschen oder nicht, und ob die Bildungsziele schon erreicht sind oder nicht. Es ist sehr wichtig, einen guten Test zu erstellen, damit man wissen kann, wie die Beherrschung von den SchülerInnen ist. Auβerdem sollen die LehrerInnen auch wissen, wie man einen guten Test erstellen kann. Der gute Test hat einige Merkmale.

Man kann durch einige Aspekten sehen, ob ein Test gut ist oder nicht. Arikunto (2012:72) hat erwähnt, die Merkmale des guten Tests sind Gültigkeit, Zuverlässigkeit, Objektivität, Praktikabilität, und ökonomischer Wert. In seinem Buch hat Nurgiyantoro (2013:108) erklärt, die Kriterien eines guten Tests sind Gültigkeit, Zuverlässigkeit, und Qualität der Items (Itemanalyse).

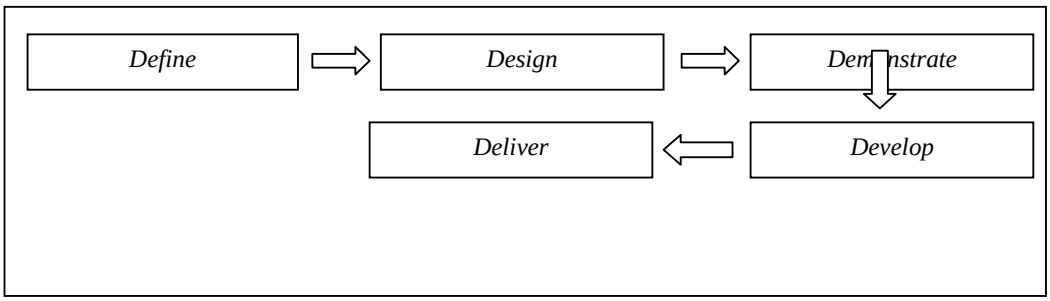
Untersuchungsmethode

Die vorliegende Untersuchung ist eine Art von der Entwicklungsuntersuchung. Die benutzte Methode in dieser Untersuchung ist die Methode von Cennamo und Kalk (2005). Sukmadinata (2009:164) und Sugiyono (2010:147) haben festgestellt, dass Untersuchung und Entwicklung ein Prozess oder Schritte sind, um ein neues Produkt zu erstellen, um ein altes Produkt zu

verbessern oder um die Effektivität des Produkts zu prüfen. Man kann durch diese Untersuchung einige Produkten erstellen.

Die erstellten Produkte sind nicht nur *hardware,* wie zum Beispiel Bücher, Module, Lernmedien, sondern auch *software,* wie zum Beispiel Computerprogramme oder Lernmodelle, usw. In dieser Untersuchung ist das erstellte Produkt ein Test für den Deutschunterricht. Um das gewünschte Produkt zu erreichen, sollte der Verfasser einige Prozedure durchführen. Die Prozedure sind den Bedarf zu analysieren, den Produktsentwurf zu bestimmen, Prototyp zu erstellen und entwickeln sowie Produktsprobe (Cennamo & Kalk, 2005:6). Die Produktsprobe zielt darauf ab, die Qualität des schon erstellten Tests zu prüfen.

In dieser Untersuchung werden fünf Prozedure angewendet. Die sogenannten Prozedure sind von Cennamo und Kalk (2005:6-7) entwickelt. Die sogenannten Prozedure sind wie folgt formuliert:



Quelle: Cennamo und Kalk (2005: 6-7)

Im *Define* Schritt hat der Verfasser das Bedürfnis und die Charakteristik der SchülerInnen identifiziert. Das Bedürfnis der SchülerInnen ist ein Test für den Deutschunterricht nach dem Curriculum 2013. Und die Charakteristik der SchülerInnen ist es, dass die SchülerInnen in der 11. Klasse des Sprachprogramms und des Programms *Lintas Minat* sind. Dann hat der Verfasser sowohl das Produkt, nämlich einen Test für den Deutschunterricht, als auch die Materialien für den Test bestimmt. Die sogenannten Materialien sind die Deutschmaterialien für die 11. Klasse des Sprachprogramms zum Thema "Alltagsleben". In diesem Schritt wurde der Verfasser auch die Aufgabentypen des Testes bestimmt, nämlich Mehrfachwahlaufgaben oder *Multiple-Choice*-Aufgaben. Es gibt fünf Auswahlmöglichkeiten. Inzwischen prüft der Test die Lesefähigkeit, Wortschatz- und Grammatikbeherrshung der SchülerInnen. Dann hat der Verfasser die Strategien geplannt, um die Qualität des Testes zu halten.

Im *Design* Schritt hat der Verfasser die Subthemen für den Deutschtest bestimmt, denn das Thema "Alltagsleben" hat drei Subthemen, nämlich "Essen und Trinken", "Kleidung" und "Wohnung".

In diesem Test wurden die Subthemen "Essen und Trinken" und "Kleidung" angewendet. Dann hat der Verfasser die Materialien für den Test gesammelt. Die Materialien sind beispielweise Texte und Bilder. Auβerdem hat der Verfasser auch den Entwurf des Testes erstellt. Im Entwurf stehen die Testspezifikationen, zum Beispiel: die Zahl der Items (80 Items), die geprüften Aspekte (Lesen, Wortschatz- und Grammatikbeherrschung), die geprüften kognitiven Bereiche und die Lösungen. Danach hat der Verfasser begonnen, die Testitems zu erstellen.

Im *Demonstrate* Schritt hat der Verfasser die Testitems erstellt. Der Entwurf wurde in den Item-Karten dargestellt. Der Verfasser hat es auch bestimmt, dass die angewendeten Materialien zu dem Bedürfnis der SchülerInnen passen. Nachdem die Item-Karten fertig erstellt wurden, wurde die Gültigkeit des Testes geprüft. Das Kriterium der Gültigkeit bedeutet, dass ein Test auch wirklich etwas überprüft, was er überprüfen soll (Albers & Bolton, 1995:22). Gültigkeit kann auch ganz einfach definiert werden als das Ausmaβ, in dem ein Test das misst, was er messen soll (ALTE, 2012:17). Es gibt viele Arten von Gültigkeit, aber jeder Test muss Inhalt-Gültigkeit haben. Um das Ziel zu erreichen, sollte eine Gültigkeitsprüfung durchgeführt werden. Die Gültigkeit des erstellten Testes wurde durch die Methode *expert judgement* geprüft. Die Validatorinnen waren vier Expertinnen von verschiedenen Bereichen, nämlich im Bereich Deutsch als Fremdsprache, der deutschen Literatur und der Evaluation.

Es gab einige Schritte, die Validatorinnen tun sollten. Dies waren wie folgt: (1) die Eignung der Grundtheorien mit dem Entwurf checken, (2) die Eignung des Entwurfs mit den Items checken, (3) die Items in der Itemskarte bewerten, ob sie relevant war oder nicht, (4) die Lösungen und die Bewertungsrubrik checken und (5) den Fragebogen ausfüllen.

Von dem Fragebogen wurde die Gültigkeit geprüft. Die Validatorinnen sollten den Fragebogen mit den Noten von eins bis vier ausfüllen. Die Zahl "1" bedeutet nicht relevant, "2" ziemlich relevant, "3" relevant und "4" sehr relevant. Die Daten des Fragebogens wurde mit der Formel Aiken's V gerechnet (Azwar, 2012:113). Die Kriterien der Gültigkeit sind in der Tabelle 1 beschrieben.

Tabelle 1 Die Kriterien der Gültigkeit

No.	Range	Kategorien
1	$0{,}80 < r \leq 1{,}00$	sehr hoch
2	$0{,}60 < r \leq 0{,}80$	hoch
3	$0{,}40 < r \leq 0{,}60$	durchschnitt
4	$0{,}20 < r \leq 0{,}40$	niedrig
5	$-1{,}00 \leq r \leq 0{,}20$	sehr niedrig

Quelle: Guilford (1956:145)

Basierend auf dem Ergebnis des Fragebogens kann es zusammengefasst, dass alle Indikatoren relevant sind. Und die Durchscnittkoeffizient ist 0,71. Der mögliche Koeffizient ist zwischen 0 – 1. Deshalb wird es gesagt, dass der erstellte Test gute Inhalt-Gültigkeit hat. Auβerdem wurden die Items auch geprüft.

Die Zahl der Items ist 80 Items, nämlich 32 für den Wortschatz, 44 für das Lesen und 4 für die Grammatik. Die Items sind angewendet, um die Lesefähigkeit, die Wortschatz- und die Grammatikbeherrsschung der SchülerInnen zu prüfen. Die Validatorinnen haben die Items gecheckt und Kommentare oder Vorschläge gegeben, damit die Items gut und relevant sind. Die Vorschläge sind wie folgt: (1) falsche Wortschatzauswahl soll verhindert werden, (2) die Indikatoren sollen mit der Grundkompetenz geeignet werden, (3) eine Antwort der Lösung ist anders als in dem Entwurf, (4) Schreibfehler sollen verbessert werden, (5) die Quellen der Texte sollen auch geschrieben werden, (6) Rechtschreibung, Struktur und Wortschatz sollen korrigiert werden, (7) die Anwendung der Symbole im Dialog soll verhindert werden, (8) Namen und Begriffe für die Sprecher sind im Dialog besser zu benutzen, (9) Aufgaben mit einem Fragesatz sollen verhindert werden, (10) die Texte sollen passend zu dem

Thema sein, (11) komplizierte Sätze sollen umformuliert werden, (12) Jeder Dialog soll einen Kontext und eine Situation haben und (13) die Bilder sollen klarer sein.

Basierend auf dem Gültigkeitsprüfungsergebnis sind 68 Aufgaben relevant, 5 relevant mit Verbesserung und 7 Aufgaben nicht relevant. Die relevanten Aufgaben können angewendet werden. Inzwischen sind 5 Aufgaben schon verbessert, damit die Aufgaben relevant sind. Und die nicht relevante Aufgaben können nicht angewendet werden. Nach der Verbesserung wurde die Testprobe durchgeführt.

Die Testprobe wurde in einer Schule durchgeführt. Die Testprobe hat zum Ziel, die Lesbarkeit des erstellten Testes zu wissen. Die Lesbarkeit umfasst die Rechtschreibung, die Formulierung der Aufgaben, die Anweisung und die Bilder. Der erstellte Test bei der Probe besteht aus 73 Items. Die Items sind die relevanten Items basierend auf dem Ergebnis der Gültigkeitsprüfung.

Die Probanden sollten nicht nur die Items beantworten, sondern auch Kommentare über die Lesbarkeit des Testes geben. Die Verbesserung des Testes wurde basierend auf den Kommentaren bzw. den Fragen nach der Testprobe, damit der Test klar werden kann. Und bei der Testprobe gab es keine zeitliche Begrenzung.

Die Probanden haben den Test gemacht und Kommentare gegeben. Eines der Probanden konnte 65 Items richtig beantworten. Das ist fast 90% der Items. Und das wenigste Ergebnis ist 41 oder nur 56% der Items. Die Probanden brauchen auch unterschiedliche Zeit, um den Test fertig zu machen. Der Test wurde mindestens 50 Minuten bis 80 Minuten von den Probanden gemacht.

Die Lesbarkeit steht auch in Verbindung mit den Texten im Test. Man muss die Zahl der Wörter und die Zeit zum Lesen wissen, die die Probanden brauchen. Dieser Prozess ist nützlich, um die Zeit des Testes zu bestimmen. Es gibt verschiedene Texte im Test. Jeder Text hat unterschiedliche Zahl der Wörter. Deswegen haben die Probanden unterschiedliche Zeit gebraucht, um den Text zu lesen und zu verstehen. Es gibt dreizehn Texte, nämlich sechs kurze Texte, fünf mittelmäβige Texte und zwei lange Texte. Die kurzen Texte haben 20 bis 47 Wörter. Die SchülerInnen brauchen 30 bis 45 Sekunden, um jeden Text zu verstehen. Inzwischen haben die mittelmäβige Texte 57 bis 82 Wörter. Um einen mittelmäβigen Text zu verstehen, brauchen die SchülerInnen eine Minute. Und die langen Texte haben 114 bis 126 Wörter. Dazu brauchen die SchülerInnen 2 Minuten, um jeden Text zu verstehen. Die Probanden haben 13 Minuten verbracht, um alle Texte fertig zu lesen und verstehen.

Die Zeit, in der die SchülerInnen ein Item antworten, wurde auch notiert. Eines der Probanden haben den Test in 50 Minuten gemacht. Er hat 13 Minuten zum Lesen verbracht. Also, er hat 30 Sekunde pro Item gebraucht. Das ist die Mindeszeit. Und dann eines der Probanden haben den in 80 Minuten fertig gemacht. Und zum Lesen hat er 13 Minuten verbracht. Das bedeutet, dass er hat durchscnittlich 55 Sekunden pro Item gebraucht. Die Zeitspanne ist von 30 bis 55 Sekunden, um ein Item zu beantworten.

Basierend auf dem Ergebnis der Probe kann es zusammengepasst, dass der Test schon gute Lesbarkeit hat. Wegen der guten Lesbarkeit braucht keine Revision durchgeführt zu werden. Der Test konnte direkt durchgeführt werden.

Im *Develop* Schritt hat der Verfasser die Items bzw. den Test basierend auf dem Ergebnis der Testprobe verbessert. Dieser Entwicklungsschritt hat gedauert, so lange die schlechte Items noch vorhanden sind. Danach wurden sie verbessert.

Im *Deliver* Schritt hat der Verfasser die Materialienquelle geschrieben und andere Sachen gemacht, die den Test fördern. Danach wurde der Test durchgeführt. Die Probanden waren 33 Personnen von der 11. Klasse des Sprachprogramms und *Lintas Minat*.

Der Test hat zum Ziel den Schwierigkeitsgrad, die Trennschärfe, die Distraktoren und die Zuverlässigkeit der Items bzw. des Testes herauszufinden. Die sogenannten Informationen wurden durch die Itemanalyse bekommen. Die schlechte Items wurden weggelassen, sodass nur gute Items angewendet werden.

Basierend auf dem Schwierigkeitsgrad und der Trennschärfe eines Items kann man wissen, ob das Item noch angewendet werden kann oder nicht. Der Schwierigkeitsgrad (*item difficulty*) bedeutet, wie schwierig oder einfach ein Item für die PrüfungsteilnehmerInnen ist (Oller, in Nurgiyantoro 2013:194). Auβerdem versteht man unter der Trennschärfe (*item discrimination*) wie scharf ein Item ist, um die Fähigkeit der SchülerInnen zu trennen (Nurgiyantoro, 2013:197).

Auβer dem Schwierigkeitsgrad und der Trennschärfe wurde die Effektivität der Distraktoren auch analysiert. Distraktoren sind die falschen Optionen bei einem *Multiple-Choice*-Item (ALTE, 2012:97). Johnson & Johnson (2002:62) haben erwähnt, *Multiple-Choice-Items consist of a direct questions or incomplete statement (stem) followed by two or more possible answers (called responses), only one of which is to be selected*. Mehrfachwahlaufgaben bestehen aus einer direkten Frage oder einer unvollständigen Aussage (*stem*), die eine oder zwei mögliche Antworten (*response*) befolgt ist, aber nur eine Antwort ist zu wählen (richtig). Die SchülerInnen sollen die richtige oder die beste Antwort wählen, denn in den Mehrfachwahlaufgaben oder *Multiple-Choice*-Aufgaben gibt es einige Auswahlmöglichkeiten, aber nur eine Antwort ist richtig und die anderen funktionieren um zu locken (Distraktoren).

Die Distraktoren sollen gut sein. Das bedeutet, dass jeder Distraktor eines Items mindestens von einem Schüler gewählt wird. Die schlechten Distraktoren wurden ersetzt, denn sie funktionieren nicht. Jedes Item hat verschiedene Zahl der Distraktoren, die ersetzt wurden. Es gibt auch Items, deren Distraktoren nicht zu ersetzen brauchen. Nicht alle Distraktoren wurden verbessert. Die Distraktoren der schlechten Items wurden nicht verbessert, denn die Items weggelassen wurden. Am Ende wurde die Zuverlässigkeit des Testes besprochen. Reliabilität ist die Zuverlässigkeit der Leistungsmessung (Albers & Bolton, 1995:25). Zuverlässigkeit bedeutet Konsistenz: Ein Test mit reliabler Bewertung führt bei jedem Einsatz zu gleichen oder ähnlichen Resultaten (ALTE, 2012:19). Anderson (in Arikunto, 2012:101) hat gemeint, dass Gültigkeit und Zuverlässigkeit in einem Test sehr wichtig ist. Es gibt einige Arten der Zuverlässigkeit. In dieser Untersuchung ist die Zuverlässigkeit des Tests mit der Formel *Alpha Cronbach* bewertet. Der Koeffizient der Zuverlässigkeit ist von 0,00 bis 1,00. Nach der Analyse ist der Koeffizient der Zuverlässigkeit *Alpha Cronbach* sehr hoch, nämlich 0,831. Nach der Itemanalyse ist die Zahl der guten Items 54 Items.

Das Entwicklungsergebnis

Das Ergebnis dieser Entwicklung ist ein Testinstrument für den Deutschunterricht nach dem Curriculum 2013 für die 11. Klasse des Sprachprogramms im zweiten Semester. Der Test prüft die Lesefähigkeit und die Grammatik- und Wortschatzbeherrschung der SchülerInnen. Der erstellte Test ist ein objektiver Test, dessen Aufgabentypen des Testes Mehrfachauswahlaufgaben oder *Multiple-Choice*-Aufgaben sind. Es gibt fünf Auswahlmöglichkeiten, vier davon funktionieren als Distraktoren. Im Test sind elf Texte vorhanden. Die Texte haben verschiedene Themen, acht davon haben das Thema "Essen und Trinken" und die andere haben das Thema "Kleidung". Sie haben auch unterschiedliche Zahl der Wörter. Die kurzen Texte haben 20 bis 47 Wörter. Die SchülerInnen brauchen 30 bis 45 Sekunden, um jeden Text zu verstehen. Inzwischen haben die mäβigen Texte 57 bis 82 Wörter. Die SchülerInnen brauchen eine Minute, um jeden Text zu verstehen. Und die langen Texte haben 114 bis 126 Wörter. Die SchülerInnen brauchen 2 Minuten, um jeden Text zu verstehen. Nachdem der Test fertig erstellt wurde, wurde die Gültigkeit des Testes geprüft.

Der estellte Test hat gute Inhalt-Gültigkeit. Um die Gültigkeit des Instruments zu prüfen, wurden zwei Schritte durchgeführt, nämlich *expert judgement* Methode und Bewertung mit der Formel Aiken's V. Der Koeffizient ist 0,71. Das bedeutet, dass der Test gültig ist. Auβerdem wurde eine Testprobe durchgeführt. Bei der Testprobe wurde auch die Zeit, in der die SchülerInnen ein Item antworten, auch notiert. Eines der Probanden haben den Test in 50 Minuten gemacht. Er hat 13 Minuten zum Lesen verbracht. Also, er hat 30 Sekunde pro Item gebraucht. Das ist die Mindeszeit. Und dann eines der Probanden haben den in 80 Minuten fertig gemacht. Und zum Lesen hat er 13 Minuten verbracht. Das bedeutet, dass er hat durchscnittlich 55 Sekunden pro Item gebraucht. Die Zeitspanne ist von 30 bis 55 Sekunden, um ein Item zu beantworten. Basierend auf dem Ergebnis der Testprobe hat der Test auch gute Lesbarkeit und gute Items. Die Zahl der Items ist 54 Items.

Die Items des Testes haben passenden Schwierigkeitsgrad und passende Trennschärfe. Die Distraktoren des Items funktionieren auch gut. Inzwischen hat der Test auch gute Zuverlässigkeit. Die Zuverlässigkeit des Testes wurde mit der Formel *Alpha Cronbach* bewertet. Der Koeffizient der Zuverlässigkeit ist 0,863. Zusammenfassend hat der Test gute Qualität.

Schlussfolgerung

Das erstellte Testinstrument ist ein Test für den Deutschunterricht nach dem Curriculum 2013 für die 11. Klasse im zweiten Semester. Im Test sind zwei Subthemen vorhanden, nämlich "Essen und Trinken" und "Kleidung". Der Test prüft die Lesefähigkeit und die Grammatik- und Wortschatzbeherrschung der SchülerInnen. Die Items wurden basierend auf der Grundkompetenz 3.3 erstellt. In der Grundkompetenz 3.3 sollen die SchülerInnen die Sprachelemente, den Struktur eines Textes und kulturelle Elemente zum Thema "Alltagsleben" verstehen. Der erstellte Test ist ein objektiver Test. Die Aufgabentypen des Testes sind Mehrfachauswahlaufgaben oder *Multiple-Choice*-Aufgaben.

Die Gültigkeit des Testes und der Items wurde durch die Methode *expert judgement* geprüft. Auβerdem wurde eine Testprobe durchgeführt und die Items

wurden auch analysiert, so dass die Items gute Qüalität hat. Die Zahl der guten Items ist 54 Items, die guten Schwierigkeitgrad, gute Trennschärfe, gute Distraktoren und gute Zuverlässigkeit haben.

Der erstellte Test haben dessen Stärke. Die Stärke sind (1) der Test ist passend zu dem Curriculum 2013, (2) der Test prüft die kognitiven Bereiche C1 – C4 (Wissen, Verstehen, Anwenden und Analysieren), (3) der Test hat gute Items basierend auf dem Ergebnis der Itemanalyse, (4) der Test hat gute Gültigkeit und Zuverlässigkeit, (5) Der Test prüft die Lesefähigkeit, Wortschatz- und Grammatikbeherrschung und (6) die Items des Testes können als Midsemestertest angewendet werden. Der Test hat nicht nur Stärke sondern auch Schwäche. Der erstellte Test hat noch Schwäche. Die Schwäche sind (1) das Thema des Testes beschränkt sich auf das Thema "Alltagsleben" und (2) Die Testprobe wurde in einer Schule durchgeführt.

Quellenverzeichnis

Albers, H-G. & Bolton, S. 1995. *Testen und Prüfen in der Grundstufe.* München: Langenscheidt.

ALTE. 2012. *Handbuch zur Entwicklung und Durchführung von Sprachtests, Zur Verwendung mit dem GER.* Frankurt am Main: telc GmbH. Anderson, W. L. & Krathwohl, D. R. 2010. *Kerangka Landasan untuk Pembelajaran, Pengajaran, dan Asesmen.* Yogyakarta: Pustaka Pelajar.

Arikunto, S. 2012. *Dasar-Dasar Evaluasi Pendidikan (Revisi Kedua).* Jakarta: Rineka Cipta.

Azwar, S. 2012. *Penyusunan Skala Psikologi (Edisi 2).* Yogyakarta: Pustaka Pelajar.

Cennamo, K & Kalk, D. 2005. *Real World Instructional Design*. Kanada: Thomson Learning, Inc.

Guilford, J. P. 1956. *Fundamental Statistics in Psychology and Education.* New York: Mc. Graw-Hill Book Co. Inc.

Johnson, D. W. & Johnson, R. T. 2002. *Meaningful Assessment (a Manageable Cooperative Process).* Boston: A Pearson Education Company.

Kementerian Pendidikan dan Kebudayaan. 2013. *Kurikulum 2013 Kompetensi Dasar SMA/MA*. Jakarta: Kementerian Pendidikan dan Kebudayaan.

Kementerian Pendidikan dan Kebudayaan.*Undang-Undang Republik Indonesia No. 20 Tahun 2003 tentang Sistem Pendidikan Nasional.* Jakarta: Kementrian Pendidikan dan Kebudayaan.

Lestari, P. P. 2014. *Validitas dan Reliabilitas Soal UAS Bahasa Jerman Kelas XI IPS SMA Negeri 7 Malang*. Diplomarbeit wurde nicht gedrückt. Malang: Staatliche Universität Malang.

Mulyasa. H.E. 2013. *Pengembangan dan Implementasi Kurikulum 2013.* Bandung: PT. Remaja Rosdakarya.

Nurgiyantoro, B. 2009. *Penilaian dalam Pengajaran Bahasa dan Sastra*. Yogyakarta: BPFE.

Nurgiyantoro, B. 2013. *Penilaian Pembelajaran Bahasa Berbasis Kompetensi.* Yogyakarta: FEB UGM.

Sugiyono. 2010. *Metode Penelitian Kuantitatif, Kualitatif, dan R&D.* Bandung: Alfabeta.

Sukmadinata, N. S. 2009. *Metode Penelitian Pendidikan*. Bandung: PT. Remaja Rosdakarya.