Erprobung eines Gruppentests zur Überprüfung des Grammatikverständnisses auf der Basis des TROG-D

Timo Lüke
Bergische Universität Wuppertal

Ute Ritterfeld und Heinrich Tröster
Technische Universität Dortmund

Korrespondenzadresse: Timo Lüke, Bergische Universität Wuppertal, Institut für Bildungsforschung, Gaußstr. 20, 42119 Wuppertal. http://orcid.org/0000-0002-2603-7341

https://doi.org/10.1026/0012-1924/a000157

Zusammenfassung

Zur Erfassung der rezeptiven grammatischen Kompetenzen liegt mit dem TROG-D (basierend auf dem englischsprachigen Test for Reception of Grammar; Bishop, 1989) ein anerkanntes Verfahren vor, das im Einzelsetting durchgeführt werden muss und so einschließlich Auswertung etwa eine halbe Stunde pro TeilnehmerIn beansprucht. In der vorliegenden Studie wurde ein auf dem TROG-D basierender Gruppentest erprobt, der den Einsatz beispielsweise in Schulklassen ermöglichen soll. Die überarbeitete Testversion wurde an 93 Zweit- und DrittklässlerInnen erprobt. Zwei Wochen später wurden alle TeilnehmerInnen mit dem TROG-D im Einzelsetting getestet. Eine ROC-Analyse (Receiver Operating Characteristic) und etablierte Screening-Gütekriterien lieferten klare Belege dafür, dass es mit dem Gruppentest gelingt, diejenigen Kinder zu identifizieren, die sich auch in der Einzeltestung mit dem TROG-D als unterdurchschnittlich erweisen. Die Ergebnisse belegen das grundsätzlich große Potential eines solchen Gruppentests und geben Ansatzpunkte zur Weiterentwicklung dieser Vorgehensweise.

Schlüsselwörter: Grammatik, Gruppentestung, Screening, Sprachentwicklung, TROG-D

https://doi.org/10.1026/0012-1924/a000157

Testing a group test for reception of Grammar based on the TROG-D

Abstract

The TROG-D (an adaption of the Test for Reception of Grammar; Bishop, 1989) is an accepted measure for the reception of grammar in German. However, it has to be conducted in an individual setting and requires about thirty minutes per participant. We adapted a group test based on the TROG-D, which allows the measurement of grammatical competencies in large groups of children, e.g. in classrooms. In the present study, we applied the new test to a sample of 93 primary students in a group setting. Two weeks later, we assessed the participants' grammatical skills using the TROG-D (individual setting). ROC (Receiver Operating Characteristic) and established screening quality criteria support the idea that the group test successfully identifies children with low grammatical performance in the TROG-D. Our results confirm the high potential of such a group test and offer insights for the advancement of this approach.

Key Words: grammar, group testing, language development, TROG

https://doi.org/10.1026/0012-1924/a000157

In der Unterrichtskommunikation kommt bildungssprachlichen Merkmalen (*Cognitive Academic Language Proficiency*; CALP, Cummins, 2008) eine entscheidende Rolle zu, denn das Verständnis der in der Schule verwendeten Instruktionssprache ist eine wesentliche Voraussetzung für den Erfolg in allen schulischen Lernbereichen. Purpura (2013) konzeptualisiert das Sprachwissen eines Kindes als Interaktion seines grammatischen und pragmatischen Wissens. Beide sind obligatorisch für ein gutes Sprachverständnis, sodass sich ein begrenztes grammatisches Wissen limitierend auf das Sprachverständnis auswirkt. Probleme im Bereich des laut- wie schriftsprachlichen Sprachverständnisses wirken sich wiederum direkt auf die Leistungen in allen Schulfächern aus (Siegmüller, 2010). Beeinträchtigungen im Sprachverständnis, insbesondere im Verständnis grammatischer Strukturen, gelten weiterhin als wichtige Warnzeichen für zukünftige und bereits bestehende Probleme in der Sprachentwicklung (Bishop, Bright, James, Bishop & van der Lely, 2000). Die meisten Kinder schließen die Entwicklung des auditiven Sprachverständnisses bereits im Einschulungsalter erfolgreich ab (Kannengieser, 2009), einige Kinder stehen aber auch im Grundschulalter noch *vor* dieser Entwicklungsaufgabe (Watermeyer, Höhle & Kauschke, 2011). Es ist also erforderlich, Probleme im Grammatikverständnis von Kindern auch noch im Grundschulalter zu identifizieren, um eine adäquate Förderung einleiten zu können. Ein Gruppentest, wie wir ihn hier vorschlagen, sollte also im Grundschulalter, insbesondere bis Klasse 4., eingesetzt werden können.

Die alltägliche Diagnostik, die viele Lehrkräfte kontinuierlich und kompetent in ihren Unterricht integrieren, sollte durch Testverfahren und Screenings ergänzt werden, deren Testgüte im Sinne evidenzbasierter Praxis (Slavin, 2002) wissenschaftlich belegt ist. Das setzt allerdings die Verfügbarkeit dieser Verfahren voraus, die im sprachdiagnostischen Bereich noch begrenzt ist (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen [IQWiG], 2009; Law, Boyle & Harris, 1998; Neugebauer & Becker-Mrotzek, 2013). Die

https://doi.org/10.1026/0012-1924/a000157

Überprüfung produktiver grammatischer Kompetenzen[1] wäre zwar wünschenswert, ist unter Gruppentestbedingungen und für den Zweck des Screenings in dieser Altersgruppe aber schwer oder gar nicht umsetzbar, da schriftsprachbasierte Testformate zu hohe Anforderungen an die Kinder stellen und lautsprachliche Antworten eine Bearbeitung in der Gruppe ausschließen. In der vorliegenden Arbeit soll daher die Möglichkeit der Überprüfung rezeptiver Grammatikkompetenzen im Gruppensetting auf Basis eines Bildauswahlverfahrens erkundet werden. Ein solcher Test ist bis dato nicht vorhanden, wird aber vielfach gefordert (Mahlau & Blumenthal, 2014; Spreer, 2013). Mit der vorliegenden Studie möchten wir daher eine Möglichkeit aufzeigen, einen solchen Test zu realisieren.

Insbesondere vor dem aktuellen Hintergrund inklusiver Schulentwicklung (vgl. Grosche & Volpe, 2013; Mahlau, Diehl, Voß & Hartke, 2011) sind im schulischen Kontext Screeningverfahren erforderlich, die die ökonomische Testung einer ganzen Klasse gestatten und der Lehrkraft die Identifikation von Kindern ermöglichen, denen eine zusätzliche Förderung und genaue Diagnostik zugutekäme. Es wäre also wünschenswert, einen Gruppentest zur Verfügung zu stellen, der unter normalen Unterrichtsbedingungen in der Klasse durchgeführt und in zufriedenstellender Weise zur Vorauswahldiagnostik genutzt werden kann.

Daraus ergibt sich der Anspruch, dass ein sprachbezogener Gruppentest (1) möglichst wenig Zeit in Anspruch nehmen, leicht durchführbar und möglichst nicht belastend für die Teilnehmenden sein sollte sowie (2) möglichst viele Kinder mit unterdurchschnittlichen Fähigkeiten als „im Screening auffällig“ und Kinder mit mindestens durchschnittlichen Fähigkeiten als „im Screening unauffällig“ klassifizieren soll. In der vorliegenden Studie wurde eine Adaption des TROG-D (Test zur Überprüfung des Grammatikverständnisses; Fox, 2013) – einem anerkannten Verfahren der Individualdiagnostik – für Gruppensettings

[1] Für einen Überblick der Möglichkeiten zur Elizitation siehe Purpura, 2012.

https://doi.org/10.1026/0012-1924/a000157

erprobt. Die Ergebnisse des Gruppentests wurden dann mit den Ergebnissen einer Individualdiagnostik mit dem TROG-D verglichen. Den Fokus legen wir dabei auf die Screening-Funktion dieses Gruppentests und berichten entsprechende Güteindizes.

Der Test zur Überprüfung des Grammatikverständnisses (TROG-D)

Der TROG-D (Fox, 2013) basiert auf dem englischsprachigen *Test for Reception of Grammar* (TROG; Bishop, 1989) und dem *Test for Reception of Grammar 2* (TROG-2; Bishop, 2003), die ursprünglich als Materialien zu Forschungszwecken konzipiert wurden, sich aber schnell als Diagnostika von Grammatikstörungen im Kindes- und Erwachsenenalter in der sprachtherapeutischen und psychologischen Praxis etablierten. Heute ist auch der TROG-D zur Diagnostik und zur Ableitung von Therapiezielen etabliert (Bub, 2007; Lohmeier, 2007; Winkler, 2007; Leibniz-Zentrum für Psychologische Information und Dokumentation [ZPID], 2007). Im Gegensatz zu anderen Testverfahren sowie den Untertests umfangreicherer Sprachentwicklungstests liegt der Fokus dabei auf der Prüfung des Verständnisses *grammatischer* und weniger auf der Prüfung semantischer Strukturen (Feldhusen, Brunner, Heinrich & Pröschel, 2007). Dabei wurden die Items des TROG-D aus dem Grammatikerwerbsmodell von Clahsen (1986) abgeleitet (Fox, 2013). Das grundsätzliche Vorgehen während der Testung mit dem TROG-D lässt sich wie folgt skizzieren: Die Testleitung spricht das jeweilige Zielwort beziehungsweise den Zielsatz vor und das Kind bzw. die oder der Erwachsene wählt aus vier Bildern auf einer vor ihm liegenden Bildkarte dasjenige aus, das der Zielstruktur entspricht. Die Bearbeitungszeit ist nicht strikt begrenzt und sowohl Wiederholungen als auch Selbstkorrekturen sind möglich. Pro grammatischer Struktur (Itemblock) werden jeweils vier solcher Items präsentiert. Itemblock K (Passiv) besteht zum Beispiel aus den Items „Das Mädchen wird vom Pferd gejagt“, „Der Elefant wird vom Jungen geschoben“, „Das Pferd wird vom Mann gejagt“ und

https://doi.org/10.1026/0012-1924/a000157

„Die Kuh wird vom Mann geschoben“ (Fox, 2013, S. 18). Ein Itemblock gilt nur dann als gelöst, wenn *alle vier* Items korrekt gelöst wurden (Bishop et al., 2000).

In der deutschen Version besteht der Test aus 21 Itemblöcken zu jeweils 4 Items. Die Distraktorbilder unterscheiden sich entweder lexikalisch oder grammatisch minimal von der Zielstruktur. Beispielsweise lauten die Distraktorbilder zur Zielstruktur „Er schiebt den Elefanten.“ (s. o.): „Sie schiebt den Elefanten.“, „Sie schieben den Elefanten.“ und „Der Elefant schiebt ihn.“ (Fox, 2013, S. 12). Die Testung wird abgebrochen, sobald fünf Itemblöcke in Folge nicht vollständig korrekt beantwortet wurden.

Die Durchführungs- und Auswertungsobjektivität des TROG-D kann bei Einhaltung der umfangreichen und klaren Vorgaben im Manual und wegen des recht simplen Zielverhaltens (Zeigen auf eines von vier Bildern) als gesichert angesehen werden. Die Interpretationsobjektivität ergibt sich aus der Bewertung der Leistung im Vergleich zu den Testergebnissen der Normierungsstichprobe. Für den TROG-D liegen Normdaten von 870 monolingual deutschsprachig aufwachsenden Kindern im Alter von 3;0 bis 10;11 aus dem Jahr 2005 vor (Fox, 2013).

Fox (2013) berichtet im Manual des TROG-D für die Normierungsstichprobe eine hohe interne Konsistenz (Cronbachs $\alpha = .90$) der 21 Itemblöcke. Auch die *Split-Half-Reliabilität* (*Odd Even*) ist mit $r = .91$ hoch. Weiterhin wurde die eindimensionale Struktur des Verfahrens mittels Hauptkomponentenanalyse bestätigt. Die Korrelation von Skalenrohwert und Faktorwert beträgt $r = .99$. Der Zusammenhang zwischen Skalenwert und Testalter lag in der Normierungsstichprobe bei $r = .82$ (Fox, 2013). Im Manual werden keine Angaben zur Retest-Reliabilität des TROG-D gemacht. Unseres Wissens liegen auch keine Studien zur Retest-Reliabilität des TROG-D oder der englischen Vorlagen vor (Fox, persönl. Mitteilung, 2013b; Gathercole, Willis, Baddeley & Emslie, 1994; Jones, Long & Finlay, 2006).

https://doi.org/10.1026/0012-1924/a000157

Fox (2013) führt die Korrelation der Testergebnisse mit den Ergebnissen des Subtests „Verstehen von Sätzen“ des *Sprachentwicklungstests für drei- bis fünfjährige Kinder* (SETK 3-5; Grimm, Aktas & Frevert, 2010) von $r = .72$ als Beleg für die konvergente Validität des TROG-D an. Sarimski (2013) fand zwischen TROG-D und den Subtests „Verstehen von Sätzen“ aus dem SETK 3-5 (Grimm et al., 2010) und dem *Heidelberger Sprachentwicklungstest* (HSET; Grimm & Schöler, 1998) sowie dem Untertest „Sätze verstehen“ aus dem *Marburger Sprachverständnistest für Kinder* (MSVK; Elben & Lohaus, 2000) Korrelationen von $r = .45$, .49 und .60.

Der TROG-D wird auch zur konvergenten (z. B. Gruber, 2012; Siegmüller, Kauschke, von Minnen & Bittner, 2011) und diskriminanten (z. B. Ricken, Fritz & Balzer, 2011; Ricken, Fritz-Stratmann & Balzer, 2013) Validierung anderer Entwicklungstests eingesetzt. Wir gehen nachfolgend deshalb davon aus, dass der TROG-D zur Erfassung grammatischer Kompetenzen grundsätzlich gut geeignet ist.

Allerdings kann dieser Einzeltest nicht für das Screening von Schulklassen verwendet werden. Die Durchführung dauert etwa 10 bis 20 Minuten, hinzukommen 5 bis 10 Minuten für die Auswertung (Bub, 2007). Ein ökonomischeres Instrument zur Erfassung der rezeptiven Sprachkompetenz auf der Basis des TROG-D würde damit einen wesentlichen Beitrag zur frühen Identifikation von grammatischen Störungen leisten.

Fragestellung

Wir untersuchen, ob ein auf dem TROG-D basierender Gruppentest des Grammatikverständnisses noch ähnlich gute Ergebnisse liefert wie die Einzeltestversion und somit grundsätzlich zum Screening in Schulklassen geeignet sein könnte. Ein solcher Gruppentest liefert zum einen Lehrkräften im (zunehmend inklusiven) Schulalltag ein höchst ökonomisches Instrument zum Screening der grammatischen Fähigkeiten ihrer Schülerinnen und Schüler und bietet sich zum anderen als Instrument für die Forschung an. Es sollen die

https://doi.org/10.1026/0012-1924/a000157

grundsätzliche Umsetzbarkeit einer Gruppenversion evaluiert und Hinweise auf die Qualität der Messungen im Vergleich zur Einzeltestung gewonnen werden. Mögliche Schwierigkeiten mit dem entworfenen Material sollen aufgedeckt und Verbesserungsmöglichkeiten erarbeitet werden. Weiterhin soll der Gruppentestentwurf auf der Grundlage der *Klassischen Testtheorie* (für einen Überblick: Moosbrugger, 2012) analysiert und mit den Ergebnissen des TROG-D verglichen werden.

Sofern sich der Gruppentest als grundsätzlich durchführbar erweist, sollen vor allem zwei Fragen beantwortet werden:

1) Wie eng ist der Zusammenhang zwischen den Ergebnissen des neu konstruierten Gruppentests und denen des Einzeltests?

2) Kann der Gruppentest sinnvoll als Screening eingesetzt werden und können damit Schülerinnen und Schüler identifiziert werden, die auch in der Einzeldiagnostik unterdurchschnittliche Leistungen zeigen?

Methode

Stichprobe

Es nahmen 93 Schülerinnen und Schüler aus einer nordrhein-westfälischen Kleinstadt an der Studie teil. Die Schule wurde zufällig aus der Schuldatenbank des Landes gezogen. Die Schülerinnen und Schüler stammten aus vier Klassen der 2. (48 %) und 3. (52 %) Jahrgangsstufe und waren bei der ersten Datenerhebung durchschnittlich 8.94 Jahre alt (SD = 0.64). Mädchen (49.5 %) und Jungen (50.5 %) waren etwa gleich stark vertreten. Zehn Teilnehmende (11 %) wuchsen mehrsprachig auf, wobei keine genaueren Informationen über die Erwerbsverläufe vorliegen. In zwei Fällen lag keine Angabe über die Ein-/Mehrsprachigkeit vor, alle anderen Teilnehmenden wuchsen einsprachig deutsch auf. Drei Kinder (3 %) hatten einen festgestellten sonderpädagogischen Förderbedarf im Förderschwerpunkt Lernen. Die von den jeweiligen Klassenlehrkräften anhand der

https://doi.org/10.1026/0012-1924/a000157

Schulnotenskala beurteilte Kompetenz im Fach Deutsch verteilte sich erwartungsgemäß (*Md* = 3, *Min-Max* = 1–5, Interquartilsabstand [*IQR]* = 1). Alle Klassen wurden seit der Einschulung von der jeweiligen Klassenlehrkraft unterrichtet.

Adaption des TROG-D

Eine Adaption des Testmaterials für den Einsatz als Gruppentest erfolgte im Rahmen einer Interventionsstudie. Dort wurde auch eine erste Überprüfung der Brauchbarkeit vorgenommen (Ritterfeld, Lüke & Eiermann, 2013). Das angepasste Instrument hatte denselben Itemstamm wie der TROG-D im Original. Es wurden also ebenfalls 21 Itemblöcke zu je vier Items verwendet, wobei der erste Itemblock zur Übung genutzt und deshalb aus den eigentlichen Analysen ausgenommen wurde. Das Testmaterial wurde wie folgt adaptiert, damit es den Anforderungen einer Gruppentestung entsprechend eingesetzt werden kann.

Testheft

Zur Erfassung der Antworten kreuzen die Teilnehmenden während des Tests in einem Testheft das richtige Bild zu jedem Item an. Die ursprünglich im DIN-A4-Format (297 mm x 210 mm) vorliegenden Bildkarten sind auf 40 % der Originalgröße verkleinert (119 mm x 84 mm), damit alle vier Items eines Blocks auf einer Seite des Testheftes Platz finden. Dadurch kann zu häufiges Umblättern in der Testzeit verhindert und der Umfang des Testmaterials geringer gehalten werden. Wie in Abbildung 1 (rechts) gezeigt, sind im Testheft pro Seite alle vier Items (z. B. K1, K2, K3, K4) des jeweiligen Itemblocks auf einmal zu sehen.

Das Deckblatt des Testheftes enthält eine verkleinerte Abbildung der Testheftseite zu Itemblock A mit zusätzlicher Nummerierung der vier Items zur Illustration (siehe Abbildung 1, links), anhand derer insbesondere die Bearbeitungsreihenfolge erläutert wird. Die Kinder markieren das richtige Bild durch Ankreuzen im Testheft.

---etwa hier Abbildung 1 einfügen.---

https://doi.org/10.1026/0012-1924/a000157

Auditive Stimuli

Um eine möglichst hohe Durchführungsobjektivität zu erreichen, wurden die auditiven Stimuli während der Untersuchung nicht von der Testleitung vorgelesen, sondern von einer CD abgespielt. Ein Gongschlag signalisierte den Teilnehmenden, dass sie die nächste Seite ihres Testheftes aufschlagen sollten. Dafür hatten Sie immer 3.5 Sekunden (Sek.) Zeit, bevor das erste Item des jeweiligen Blocks zu hören war. Dann sollten die Teilnehmenden das richtige Bild identifizieren und ankreuzen, bevor dann automatisch das zweite, dritte und vierte Item eingespielt wurden.

Die Bearbeitungszeit pro Item wurde nach mehrmaliger Pilotierung wie folgt festgelegt: In den ersten vier Itemblöcken (A/B/C/D) standen jeweils 2 Sek. pro Item zur Verfügung. Für die restlichen Items erhöhte sich die Bearbeitungszeit alle drei Itemblöcke um je eine halbe Sek.: In den Itemblöcken E/F/G standen 2.5 Sek. je Item zur Verfügung. Die ansteigende Bearbeitungszeit ist erforderlich, da die auditiven Stimuli nicht nur grammatisch komplexer werden, sondern auch zeitlich umfangreicher (z. B. „lang“ [C2] vs. „Während das Mädchen reitet, isst es einen Apfel.“ [P2]).

Die Begrenzung der Bearbeitungszeit wurde vorgenommen, um das Risiko von Störungen im Testablauf durch zu große Wartezeiten bei einzelnen Schülerinnen und Schülern zu minimieren.

Durchführung

Zunächst wurde der so konzipierte Gruppentest klassenweise mit allen Teilnehmenden durchgeführt.

Die Gruppentestungen erfolgten in den jeweiligen Klassenräumen durch den Erstautor im Beisein der jeweiligen Klassenlehrkraft, die zeitgleich die Klassenliste ausfüllte und sich nicht am Geschehen im Klassenraum beteiligte. Die Präsentation der auditiven Stimuli erfolgte über einen CD-Player, der auf 65dB eingestellt wurde, sodass die auditiven Stimuli

https://doi.org/10.1026/0012-1924/a000157

für alle Schülerinnen und Schüler gut hörbar waren. Die Bearbeitungszeit für jedes Item war durch die Dauer der jeweiligen Ruhephase bis zum Signal bzw. nächsten Item auf der abgespielten CD klar begrenzt. So wurden identische Bearbeitungszeiten für alle Kinder sichergestellt.

Der Umgang mit dem Aufgabenformat und dem Testheft wurde mit allen Kindern anhand des Deckblattes und eines Übungsblockes trainiert. In diesem Rahmen wurden alle Fragen der Kinder beantwortet. Während der Bearbeitungszeit wurde die Testung nicht unterbrochen. Da die Kinder während der Bearbeitung keine Fragen mehr hatten, war dies aber auch nicht notwendig. Alle Gruppentestungen verliefen störungsfrei und ohne besondere Vorkommnisse.

Die Auswertung des Gruppentests wurde auf Itemebene und nicht wie beim TROG-D auf Ebene der Itemblöcke vorgenommen. Da die Kinder die Aufgaben unter Zeitdruck bearbeiteten, kann das nicht vollständig korrekte Bearbeiten eines Itemblocks (z. B. das vierte Item des Blocks falsch) nicht mehr nur auf das Nichtbeherrschen der jeweiligen grammatischen Struktur zurückgeführt werden, sondern auch auf die Bearbeitungsgeschwindigkeit. Da es ohnehin nicht Zweck des Screenings ist, differenzierte Aussagen über einzelne grammatische Zielstrukturen zu treffen oder spezifische Therapieziele abzuleiten, werteten wir stattdessen die jeweilige Anzahl korrekt bearbeiteter Aufgaben aus. Die Information darüber, ob ein Kind zwei oder drei Items eines Blocks korrekt bearbeitet hat, geht so nicht in einer pauschalen Bewertung des Itemblocks als „falsch“ unter (d. h. nicht alle Items des Blocks korrekt), sondern fließt in einen Gesamtscore ein.

Zwei Wochen später wurden alle Teilnehmenden unter optimalen Bedingungen mit dem TROG-D im Einzelsetting getestet. Erfahrungen in den Pilotierungen hatten gezeigt, dass die Teilnehmenden sich im Anschluss an die Gruppentestung – vermutlich wegen der kurzen

https://doi.org/10.1026/0012-1924/a000157

Bearbeitungszeiten pro Item – nur an wenige Inhalte des Tests erinnern konnten. Die Einzeltestungen wurden an zwei Tagen vom Erstautor und drei geschulten Studierenden der Sonderpädagogik bzw. Klinischen Linguistik höherer Semester durchgeführt. Sie fanden in ruhigen Fachräumen der Schule statt. Nur zwei Kinder erreichten das Abbruchkriterium des TROG-D (fünf fehlerhafte Itemblöcke in Folge). Die Einzeltestungen verliefen ebenfalls störungsfrei und ohne besondere Vorkommnisse.

Zehn Prozent der Testhefte (Gruppentestung) und Protokollbogen (Einzeltestung) wurden zufällig ausgewählt und doppelt kodiert. Aufgrund des einfachen Formats und der klaren Zuordnung der Antworten betrug die Übereinstimmung 100 % und es wurde auf weitere Doppelkodierungen verzichtet. Für die Gruppentestung wurden nicht eindeutige und fehlende Antworten mitkodiert. Bei den Einzeltestungen traten diese durch Einhaltung der Anweisungen im Testmanual nicht auf.

Die Auswertung der Daten erfolgte mit IBM SPSS Statistics 21.0.0.1. Zur Berechnung einiger Güteindizes wurde zusätzlich die „Auswertungshilfe zur Berechnung von Testkennwerten von Screeningverfahren“ von Lenhard und Marx (2010) genutzt. Für die Imputation wurde HOTDECK für SPSS von Myers (2011b) verwendet.

Umgang mit fehlenden Werten

In einigen Fällen traten im Gruppentest nicht eindeutige Antworten auf. Darunter fielen 1) große Markierungen, die zwei oder mehr Antwortmöglichkeiten schnitten, 2) Markierungen, die außerhalb der eigentlichen Bildflächen lagen und nicht eindeutig einer Antwortmöglichkeit zuzuordnen waren und 3) nicht klar nachvollziehbare Selbstkorrekturen (mehrere Kreuze oder Pfeile). Diese wurden als fehlende Werte klassifiziert und werden nachfolgend „fehlend“ genannt, obwohl de facto eine Antwort gegeben wurde. Insgesamt lagen 65 fehlende Werte vor (< 1 % des Gesamtdatensatzes). Die meisten Variablen wiesen nur einen einzelnen (1 %) – wenige Variablen bis zu 4 % fehlende Werte auf.

https://doi.org/10.1026/0012-1924/a000157

Es ergaben sich keine häufiger auftretenden Muster fehlender Werte, sodass keine systematischen Fehler in den Auswertungen zu erwarten sind. Statistisch abgesichert werden kann die unsystematische – also zufällige – Verteilung der fehlenden Werte beispielsweise anhand des MCAR-Tests nach Little und Rubin (1987) sowie Sijtsma und van der Ark (2003). Für den vorliegenden Datensatz ist die Nullhypothese (d.h. dass die fehlenden Werte zufällig verteilt sind) nicht zu verwerfen, $\chi^2 = 2\,524.48$, $df = 2\,416$, $p > .05$. Es ist daher wahrscheinlich, dass die nicht eindeutigen Antworten im Gruppentest zufällig fehlen und mit der für dichotome Antworten geeigneten *Hot-Deck-Imputation* ersetzt werden können. Dazu werden zunächst alle Fälle anhand möglichst vollständiger und diskreter Variablen (Deck-Variablen) in Decks sortiert. Anschließend wird jedem Fall mit fehlendem Wert (Empfänger), der Wert eines vollständigen Falls (Spenders) zugeordnet, der zufällig (ohne Zurücklegen für mögliche weitere Empfänger = Einfachimputation) aus dem jeweiligen Deck gezogen wurde (Myers, 2011a).

Für den vorliegenden Datensatz wurden die Variablen Schuljahr und Geschlecht als Deck-Variablen benutzt. Es konnten alle fehlenden Werte imputiert werden, sodass für die nachfolgenden Analysen ein vollständiger Datensatz vorliegt.

Ergebnisse

Einzeltest

Da die Stichprobe eher im oberen Bereich der Altersspanne des TROG-D liegt, zeigen sich bei allen Messungen linksschiefe Verteilungen der Testleistungen. Mit durchschnittlich 16.41 ($SD = 2.63$) von 21 möglichen, korrekt gelösten Itemblöcken zeigen die Teilnehmenden insgesamt altersgemäße Leistungen: Der mittlere T-Wert beträgt 50.46 ($SD = 10.45$) und weicht damit nicht signifikant von der Normverteilung mit $M = 50$ und $SD = 10$ ab, $t(92) = 0.427$, $p > .05$.

https://doi.org/10.1026/0012-1924/a000157

Bewertet man die Testleistungen der Teilnehmenden anhand der üblichen klinischen Kriterien, so zeigen 15 % überdurchschnittliche (T-Wert > 60), 72 % durchschnittliche (40 < T-Wert ≤ 60), 10 % unterdurchschnittliche (30 < T-Wert ≤ 40) und 3 % weit unterdurchschnittliche (T-Wert ≤ 30) Leistungen.

Fünf der 21 Itemblöcke des Einzeltests konnten von allen Teilnehmenden gelöst werden. Die Reliabilität der übrigen 16 Itemblöcke wurde wegen der dichotomen Struktur (richtig/falsch) als interne Konsistenz nach der *Kuder-Richardson-Formel* (KR-20, Kuder & Richardson, 1937) berechnet. Das Maß entspricht dem bekannteren Cronbachs α für dichotome Daten und liegt mit α = .72 erwartungsgemäß niedriger als in der Normierung (Fox, 2013). In Tabelle 1 sind die Itemkennwerte aus der Normierungsstichprobe und der vorliegenden Studie dargestellt. Die Lösungsraten fallen (aufgrund des altersbedingten Deckeneffekts) höher und die Trennschärfen niedriger aus. Nach den Ergebnissen von Schmitz und Fox (2007) ist der Deckeneffekt in dieser Altersgruppe aber erwartungskonform.

---etwa hier Tabelle 1 einfügen.---

Gruppentest

Die Auswertung des Gruppentests wurde dahingehend angepasst, dass nicht mehr nur jeder komplett gelöste Itemblock (alle vier Items) als dichotome Richtig-Falsch-Antwort gezählt wurde, sondern die Anzahl der gelösten Items innerhalb des Itemblocks. Pro Itemblock ergibt sich dementsprechend ein Score zwischen 0 und 4. Der Testrohwert wurde als Summe aller korrekt beantworteten Items berechnet. Dadurch erhöhte sich der erreichbare Testrohwert von 21 auf 80 (eigentlich 84, die ersten vier Items fielen aber weg, weil sie als Beispiele dienten). Eine Ratekorrektur wurde nicht vorgenommen, da dies nur sinnvoll ist, wenn „alle Antworten die gleiche Attraktivität oder Schwierigkeit aufweisen“ (Pospeschill,

https://doi.org/10.1026/0012-1924/a000157

2010, S. 87), was im TROG-D augenscheinlich nicht der Fall ist und auch die Anzahl der Aufgaben mit 80 recht hoch ist.

In Tabelle 2 sind dementsprechend der Itemscore (einschließlich Standardabweichung und Vertrauensintervall), die mittlere Itemschwierigkeit, die *part-whole*-korrigierte Trennschärfe und Cronbachs α des Testwerts ohne den betreffenden Itemblock dargestellt. Um einen Vergleich zu den Ergebnissen der TROG-D-Einzeltestungen ziehen zu können, wurde zusätzlich auch die Schwierigkeit der Itemblöcke wie im TROG-D berechnet.

---etwa hier Tabelle 2 einfügen.---

Wie schon beim Einzeltest zeigen sich insgesamt sehr hohe Lösungsraten. So haben nur 8 von 20 Itemblöcken eine Schwierigkeit unter .80 und 2 von 20 eine Schwierigkeit unter .50. Die *part-whole*-korrigierte Trennschärfe liegt mit Werten zwischen .22 und .61 (M = .40, SD = .11) im mittleren Bereich. Die interne Konsistenz des Gruppentests ist mit Cronbachs α = .82 zufriedenstellend.

Im Durchschnitt erreichten die Teilnehmenden einen Testrohwert von 62.86 (SD = 8.51). Die Verteilung der Testrohwerte in der Stichprobe ist erwartungsgemäß linksschief (*Schiefe* = -1.05). Die Drittklässlerinnen und -klässler zeigen durchschnittlich höhere Testrohwerte (M = 66.96, SD = 5.98) als die Zweitklässlerinnen und -klässler (M = 58.49, SD = 8.69, $t(91) = 5.51$, $p < .001$). Die nachfolgenden Analysen sollten über die Klassenstufen hinweg durchgeführt werden. Um *altersbereinigte* Testrohwerte zu erhalten, wurden die Testrohwerte innerhalb der beiden Klassenstufen z-transformiert.

Der Zusammenhang zwischen der Leistung im Gruppentest und der Lehrerkraftangabe über die Leistungen im Fach Deutsch wurde wegen der erhöhten Anzahl von Rangbindungen in beiden Variablen mit dem konservativeren *Kendall-Tau-b* (Kendall, 1962) berechnet. Für die vorliegende Stichprobe lässt sich dennoch ein mittlerer Zusammenhang von $r_\tau = -.45$, $p < .001$ nachweisen. Die unkorrigierte *Pearson-Korrelation* zwischen dem Ergebnis im TROG-

https://doi.org/10.1026/0012-1924/a000157

D-Einzeltest und dem neu konzipierten Gruppentest beträgt $r = .52$, $p < .01$, die doppelt minderungskorrigierte Korrelation $r = .68$.

Gruppentest als Screening

Der Gruppentest soll insbesondere diejenigen Schülerinnen und Schüler identifizieren, die im Einzeltest unterdurchschnittliche Leistungen ($T \leq 40$) zeigen. Um einen effizienten Einsatz des Gruppentests als Screening zu ermöglichen, müssen altersspezifische Cut-Off-Werte definiert werden, die ein möglichst ausgewogenes Verhältnis zwischen Sensitivität und Spezifität herstellen. Es soll also jeweils ein Grenzwert definiert werden, der im Screening zu einem „auffälligen“ vs. „unauffälligen“ Befund führt und dabei möglichst wenige Fehler (Falsch-Positive und Falsch-Negative) verursacht. Zur datenbasierten Festlegung solcher Cut-Off-Werte wird die *Receiver Operating Characteristic*-Analyse eingesetzt (siehe beispielsweise Vossler-Thies, Stevens, Engel & Licha, 2013; Zimmermann et al., 2013). Ein Überblick zur ROC-Analyse findet sich unter anderem bei Tröster (2009). Anschaulich wird die Methode durch die Abbildung der ROC-Kurve (Abbildung 2). Dabei wird für jeden möglichen Cut-Off-Wert des Klassifikators (Punktzahl im Screening) die Quote korrekter, positiver Screening-Befunde (Sensitivität) gegen die Fehlalarm-Quote (1 – Spezifität) geplottet. Daraus ergibt sich eine aufsteigende Kurve im oberen linken Teil des Diagramms über der Diagonalen, welche die Differenzierungsfähigkeit einer zufälligen Zuordnung zu positivem vs. negativem Screening-Befund veranschaulicht (Goldhammer & Hartig, 2012; Tröster, 2009).

---etwa hier Abbildung 2 einfügen.---

Die *Area Under the Curve* (AUC), die als Güteindex des Klassifikators interpretiert werden kann, ist groß (AUC = .86, $p < .001$, 95 % CI [.76, .96]), was für die grundsätzlich gute Differenzierungsfähigkeit des Gruppenscreenings spricht.

https://doi.org/10.1026/0012-1924/a000157

Ein Punkt auf der ROC-Kurve, der einen großen Abstand von der Diagonalen (Zufall) hat, ist durch ein gutes Verhältnis zwischen Sensitivität (SN) und Spezifität (SP) des Klassifikators gekennzeichnet (Goldhammer & Hartig, 2012). Dies drückt sich auch in der Tatsache aus, dass der *Youden-Index* (YI; definiert als YI = SN + SP - 1) hier am höchsten ist. Unter der Annahme, dass beide Fehlerarten (falsch-negative und falsch-positive Screeningbefunde) mit derselben Dringlichkeit verhindert werden sollen, wäre der Punkt mit dem höchsten YI also der ideale Cut-Off-Wert (Tröster, 2009).

Aufgrund der deutlichen Aussparung in der ROC-Kurve im oberen linken Quadranten, ergeben sich zwei mögliche Cut-Off-Werte, die in Abbildung 2 nummeriert sind und nachfolgend bezüglich ihrer Güte und Nutzbarkeit im Sinne eines Screenings verglichen werden. Punkt 1 (z = -0.90) entspricht dabei einem Prozentrang von 18 % und Punkt 2 (z = -0.25) einem Prozentrang von 33 %.

In Tabelle 3 sind zusätzlich die üblichen Deskriptoren und Güteindizes (Marx & Lenhard, 2011; Tröster, 2009) für die beiden möglichen Cut-Off-Werte dargestellt. Darunter die jeweilige Anzahl falsch-negativer (FN), richtig-positiver (RP), falsch-positiver (FP) und richtig-negativer (RP) Befunde in unserer Studie. Die Selektionsquote (SQ) gibt den Anteil positiver Screening-Befunde (RP + FP), die Trefferquote (TQ) die Quote der korrekten Screening-Befunde (RP + RN) bezogen auf die Gesamtgruppe an. Die Sensitivität ist definiert als Quote korrekter, positiver Screening-Befunde: SN = RP/(FN + RP). Die Spezifität ist definiert als Quote korrekter, negativer Screening-Befunde: SP = RN/(FP + RN). Die positive Korrektheit gibt das Verhältnis zwischen den korrekt positiven Screening-Befunden und allen positiven Screening-Befunden an: PK = RP/(RP + FP). Der *Relative Anstieg der Trefferquote gegenüber der Zufallstrefferquote* (RATZ; Marx, 1992) dokumentiert, in welchem Ausmaß sich die Trefferquote durch die Anwendung eines Screenings im Vergleich zur zufälligen Klassifikation verbessert.

https://doi.org/10.1026/0012-1924/a000157

---etwa hier Tabelle 3 einfügen.---

Mit dem strengeren (im Sinne zu erreichender Rohwertpunkte) Cut-Off-Wert 2 werden alle bis auf einen unterdurchschnittlichen Teilnehmenden erkannt. Dementsprechend hoch ist die Sensitivität (SN = 92 %). Jedoch basiert die hohe Aufdeckungsrate auf einer recht hohen Selektionsquote (SQ = 33 %) und geht deshalb mit einer ebenfalls hohen Anzahl falsch-positiver Befunde (n = 20) und einer entsprechend geringen Spezifität (SP = 75 %) einher. Mit Cut-Off-Wert 1 wird bei einer deutlich niedrigeren Selektionsquote von 18 % immer noch der größte Teil der unterdurchschnittlichen Kinder identifiziert (SN = 75 %), gleichzeitig aber sehr viel spezifischere Befunde erzeugt (SP = 90 %). Die positive Korrektheit der Klassifikation liegt bei 53 % und die Trefferquote erreicht 88 %.

Der YI ist für beide potentiellen Cut-Off-Werte hoch (Punkt 1: .65; Punkt 2: .67) und auch der RATZ ist mit .69 beziehungsweise .88 in beiden Fällen als sehr gut zu bewerten (Marx & Lenhard, 2011). Auf Ebene der Testrohwerte liegt Cut-Off-Wert 1 für Zweitklässlerinnen und -klässler bei 50 und für Drittklässlerinnen und -klässler bei 61 von 80 möglichen Punkten.

Diskussion

Ziel der vorliegenden Arbeit war es zu überprüfen, ob ein bereits etablierter Einzeltest zur Überprüfung der rezeptiven grammatischen Kompetenzen von Schülerinnen und Schülern für Gruppentestungen angepasst werden kann. Insbesondere sollte analysiert werden, ob dieser Gruppentest als Screening für die Schulpraxis von Nutzen sein könnte. Insgesamt belegt die Studie das Potential eines solchen Gruppentests.

Die Objektivität des Gruppentests wurde durch eine standardisierte Testinstruktion und klare Auswertungsregeln sichergestellt. Der Verlauf der Testungen im Rahmen unserer Studie spricht für die Klarheit des Vorgehens, da im Anschluss an die Instruktionen keine Fragen mehr offen blieben, die nicht eigentlich schon durch die Instruktion beantwortet

https://doi.org/10.1026/0012-1924/a000157

waren. Möglicherweise sollte noch stärker betont werden, wie wichtig *eindeutige* Antwortmarkierungen für den Test sind und wie diese aussehen sollten. Zwar war der Anteil dieser unklaren Antworten insgesamt gering, es wäre aber sicher vorteilhaft, sie vollständig auszuschließen.

Die Präsentation der auditiven Stimuli von einer CD wiederum dürfte einen Zugewinn bei der Durchführungsobjektivität gegenüber dem TROG-D bedeuten. Ob die Durchführung des Gruppentests zur Anwendung mit jüngeren Kindern angepasst werden muss, z. B. durch Verlängerung der Bearbeitungszeiten oder altersangepasste Instruktionen, müssen zukünftige Studien zeigen. Die Testerfahrungen *dieser* Studie zeigen, dass die Bearbeitungszeiten für fast alle Teilnehmenden angemessen waren. Die Objektivität des Gruppentests dürfte auf ähnlich hohem Niveau liegen wie die des Einzeltests. Die Schwierigkeit des Tests liegt im Vergleich zur Normstichprobe deutlich höher, was dadurch zu erklären ist, dass das Alter der Stichprobe im oberen Bereich der TROG-D-Normierung liegt. Lediglich Itemblock S (Relativsätze – Pronomen im Akkusativ/Dativ) fällt durch vergleichbare Schwierigkeit auf. Wir konnten keine befriedigende Erklärung dafür finden. Häufige fehlende Werte oder ungewöhnliche Antwortmuster lagen in keinem Fall vor und es gab auch keine besonderen Ereignisse während der Durchführung (z. B. Störungen).

Mit Cronbachs $\alpha = .82$ verfügt der Gruppentest über eine zufriedenstellende Reliabilität. Sie liegt niedriger als die des Einzeltests in der Normierungsstichprobe ($\alpha = .90$), aber höher als die des Einzeltests in der hier untersuchten Stichprobe ($\alpha = .72$).

Es ist nicht auszuschließen, dass durch die Darbietungsweise der Aufgaben im Gruppentest wie auch im Einzeltest nicht ausschließlich das Grammatikverständnis, sondern auch andere Teilkompetenzen wie das Hörverstehen notwendig sind, um die Aufgaben erfolgreich zu bearbeiten. Der Zusammenhang ($r = .68$) zwischen den Einzel- und Gruppentestleistungen legt nahe, dass in den beiden Testbedingungen nicht ausschließlich

https://doi.org/10.1026/0012-1924/a000157

dieselbe Kompetenz erfasst wird. Aufgrund der Speedbedingung könnte im Gruppentest, verglichen mit dem Einzeltest, auch der Einfluss anderer Fähigkeiten, wie etwa der Konzentrationsfähigkeit oder der allgemeinen kognitiven Verarbeitungsgeschwindigkeit zunehmen. Hinzukommt, dass nicht zwischen Fehlern aufgrund der begrenzten Zeit und Fehlern aufgrund des grammatischen Verständnisses unterschieden wird. Die gewählte Variante erscheint als die beste Alternative: „Auf jeden Fall stellt die Verquickung von Qualität und Geschwindigkeit bei ‚gespeedeten' Leistungstests ein Problem dar, für dessen Lösung es zwar einige Ansätze in der Testtheorie gibt, von denen aber keiner ganz befriedigend ist. Günstiger ist es, die *Bearbeitungszeit pro Aufgabe* zu begrenzen. Hier hat jede Person dieselben Bedingungen für jede Aufgabe und es lassen sich die meisten Testmodelle problemlos anwenden" (Rost, 2004, S. 44).

Die mittlere Korrelation zwischen der Leistung im Gruppentest und der von den Lehrkräften angegebenen Note im Fach Deutsch deutet konvergente Konstruktvalidität an. Da die Noten hier jedoch nur bedingt für die Validitätsprüfung geeignet sind, müssen in weiteren Studien zusätzliche Verfahren zum Einsatz kommen.

Der Gruppentest erweist sich als deutlich ökonomischer als der Einzeltest: In keiner der vier Klassen dauerte die Testung inklusive Begrüßung, Erklärungen, Beantworten von Fragen, Ein- und Austeilen der Testhefte länger als 25 Minuten. Damit nimmt die Gruppentestung nicht viel mehr Zeit in Anspruch als eine Einzeltestung.

Noch verbesserungswürdig ist der Materialaufwand des Gruppentests. Für die Testhefte, die in dieser Studie eingesetzt wurden, entstanden Kosten, die mit denen anderer Testverfahren (z. B. ELFE 1-6, DEMAT etc.) vergleichbar sind. Bei einer höheren Auflage wären diese Kosten noch geringer. Sie könnten auch durch doppelseitig bedruckte oder wiederverwendbare Testhefte und ähnliche Veränderungen weiter verringert werden. Eine

https://doi.org/10.1026/0012-1924/a000157

weitere Option, den Test noch kostengünstiger zu gestalten, wäre der Verzicht auf farbige Bilder. In zukünftigen Projekten könnte also auch diese Option erprobt werden.

Gruppentest als Screening

Die doppelt minderungskorrigierte Korrelation der Testleistungen im Gruppen- (Cronbachs $\alpha = .82$) und im Einzeltest (Cronbachs $\alpha = .72$) liegt bei $r = .68$. Dies legt nahe, dass ein solcher Gruppentest nicht exakt dasselbe misst, wie der Einzeltest und daher nicht für die Individualdiagnostik geeignet ist. Wie eingangs beschrieben wurde, sollte überprüft werden, ob der Gruppentest, als ökonomischere Variante des Einzeltests, Potential für die Verwendung als Vorauswahldiagnostik oder Screening hat. So könnte er eingesetzt werden, wenn im Vorfeld der eigentlichen Datenerhebungen die rezeptiven grammatischen Kompetenzen als Voraussetzung für andere Tests überprüft werden sollen oder Kinder mit höherem Risiko im grammatischen Bereich identifiziert werden sollen. Neben dem Einsatz eines solchen Gruppentests in Forschungsprojekten ist vor allem der Einsatz als Screening in der Schulpraxis angedacht.

Um zu überprüfen, ob der Gruppentest für die beschriebene Vorauswahldiagnostik bzw. das Screening geeignet ist, wurde post hoc überprüft, inwiefern ein Cut-Off-Wert für die Leistung im Gruppentest gefunden werden kann, der die spätere *klinische Auffälligkeit* im TROG-D (unterdurchschnittliches Ergebnis) möglichst effizient vorhersagt. Dazu wurden eine ROC-Analyse durchgeführt, die üblichen Güteindizes zur Beurteilung von Screenings berechnet und diese für zwei denkbare Cut-Off-Werte verglichen. Die ROC-Analyse, die selbst ein Maß für die Güte eines Klassifikators ist, ergab eine beachtliche AUC von .86. Die genaueren Analysen unterstützen die Annahme, dass der Gruppentest insgesamt trennscharf zwischen unterdurchschnittlichen und mindestens durchschnittlichen Teilnehmenden unterscheiden kann.

https://doi.org/10.1026/0012-1924/a000157

Cut-Off-Wert 2 zeigte zwar eine hohe Sensitivität von 92 %, diese basiert aber auf der hohen Selektionsquote von 33 %. Der Einsatz dieses Cut-Off-Wertes hätte in einer durchschnittlichen Schulklasse mit 27 Schülerinnen und Schülern also einen positiven Screening-Befund bei neun Schülerinnen und Schülern zur Folge.

Wenn als empfohlene Konsequenz aus dem positiven Screening-Befund eine genauere Diagnostik (z. B. Analyse anderer Sprachdaten) folgen sollte, müssten in jeder Klasse etwa neun Kinder genauer diagnostiziert werden. Diese Selektionsquote scheint vor dem Hintergrund der Zumutbarkeit (zeitlicher Aufwand und unberechtigte Sorge) für die betroffenen Schülerinnen und Schüler und Lehrkräfte sowie die Akzeptanz bei Eltern und Lehrkräften nicht haltbar. Dabei gelten auf individueller Ebene dieselben starken Einschränkungen bei der Interpretation des Screening-Ergebnisses wie bei jedem anderen Screening auch: Ein positiver Screening-Befund sollte lediglich dazu genutzt werden, eine genauere und individuelle Diagnostik in Betracht zu ziehen. Auf keinen Fall kann und soll ein Screening eine differenzierte Diagnostik ersetzen.

Cut-Off-Wert 1 erreicht bei einer Selektionsquote von 18 % immer noch eine Sensitivität von 75 % und eine Spezifität von 90 %. Vor allem liegt die positive Korrektheit mit 53 % aber auch deutlich über der von Cut-Off-Wert 2. Die positive Korrektheit der Screening-Befunde ist uns angesichts der Konsequenzen im Schuleinsatz besonders wichtig. Weitere Screening-Gütekriterien wie YI und RATZ deuten auf eine ausgewogene Klassifikation hin. In der 27 Schülerinnen und Schüler großen Beispielklasse würden aus diesem Cut-Off-Wert durchschnittlich knapp fünf positive Screening-Befunde resultieren. Unter Annahme der hier gefundenen Grundquote wäre pro Klasse mit durchschnittlich 3.5 Kindern zu rechnen, die im TROG-D tatsächlich unterdurchschnittliche Leistungen zeigen würden. In der Mehrheit der Fälle mit positivem Screening-Befund wäre eine weitere Diagnostik also tatsächlich angebracht. Bei der Schadensabwägung (Tröster, 2009) und dem

https://doi.org/10.1026/0012-1924/a000157

Einbezug des ökonomischen Einsatzes im Schulalltag, kommen wir zu dem Schluss, dass Cut-Off-Wert 1 für den vorgesehenen Zweck eines Klassenscreenings besser geeignet ist.

Die dargestellten Auswertungen basieren auf der Anzahl korrekt gelöster Items und nicht wie im TROG-D auf der Anzahl korrekt gelöster Itemblöcke. Wir haben die Güte der Klassifikation (AUC) und der Cut-Off-Werte sowie die interne Konsistenz des Gruppenscreenings auf Basis der Itemblöcke wiederholt. Alle Kennwerte verändern sich dadurch nur marginal.

Insgesamt gelingt die Erkennung der im TROG-D unterdurchschnittlichen Teilnehmenden mit dem Gruppentest sehr gut. Ein solcher Gruppentest hat daher erhebliches Potential für den Einsatz sowohl in der Forschung als auch als Screening in der Schulpraxis. Die Ergebnisse sind vor dem Hintergrund der benannten Probleme als vorsichtiger Hinweis darauf zu bewerten, dass eine Gruppentestversion umsetzbar und diagnostisch nützlich sein könnte. Hier ist insbesondere die kleine und auf Zweit- und Drittklässlerinnen und -klässler begrenzte Stichprobe zu reflektieren. Inwiefern der Gruppentest auch in anderen Altersgruppenfunktionieren würde, kann zum jetzigen Zeitpunkt nicht beurteilt werden, wobei besonders der Einsatz in der ersten Klasse oder gar im späten Vorschulalter interessant wäre. In Folgestudien, insbesondere bei Ausweitungen des Altersbereichs hin zu jüngeren Kindern, müsste die Verständlichkeit der Instruktion möglicherweise durch zusätzliche Übungsitems stärker abgesichert und die Bearbeitungszeiten der Items überprüft werden.

In zukünftigen Studien muss vor allem eine größere und repräsentative Stichprobe betrachtet werden. Diese sollte dann auch auf andere Altersgruppen ausgeweitet werden. Im Rahmen einer größeren Erhebung wäre dann auch eine Evaluation unter dem Paradigma der *Item-Response-Theorie* (für einen Überblick siehe Eid & Schmidt, 2014; Rost, 2004) sinnvoll. In jedem Fall sollten zukünftige Studien auch die Abgrenzung von anderen

https://doi.org/10.1026/0012-1924/a000157

Teilkompetenzen wie zum Beispiel Hörverstehen, Konzentrationsfähigkeit, Verarbeitungsgeschwindigkeit usw. berücksichtigen.

In unserer Untersuchung stehen lediglich der TROG-D und das Lehrkrafturteil als konvergente Validitätskriterien zur Verfügung. In zukünftigen Untersuchungen sollten zusätzliche Daten erhoben werden, um über einen differenzierteren, diagnostischen Gold-Standard zu verfügen. Weitere Desiderate ergeben sich aus der Tatsache, dass sowohl für den Gruppentest als auch für den Einzeltest (das Originalverfahren) bis dato weder Studien zur Retest-Reliabilität noch zur Messinvarianz bei mehrsprachig aufwachsenden Kindern vorliegen.

Gemessen an den Qualitätsmerkmalen für die Güte von Sprachstandsverfahren von Becker-Mrotzek et al. (2013) könnte ein solcher Gruppentest ein für die Vorauswahldiagnostik objektives, reliables und valides Instrument sein, bei dem sich die notwendige Weiterentwicklungs- und Evaluationsarbeit auszahlen könnte.

https://doi.org/10.1026/0012-1924/a000157

Literatur

Becker-Mrotzek, M., Ehlich, K., Füssenich, I., Günther, H., Hasselhorn, M., Hopf, M. et al. (2013). *Qualitätsmerkmale für Sprachstandsverfahren im Elementarbereich. Ein Bewertungsrahmen für fundierte Sprachdiagnostik in der Kita.* Köln: Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache. Verfügbar unter http://www.mercator-institut-sprachfoerderung.de/fileadmin/user_upload/Institut_Sprachfoerderung/Mercator-Institut_Qualitaetsmerkmale_Sprachdiagnostik_Kita_Web.pdf

Bishop, D. V. M. (1989). *TROG. Test for Reception of Grammar*. Manchester: University of Manchester.

Bishop, D. V. M. (2003). *TROG-2. Test for Reception of Grammar 2*. London: Pearson.

Bishop, D. V. M., Bright, P., James, C., Bishop, S. J. & van der Lely, H. K. J. (2000). Grammatical SLI: A distinct subtype of developmental language impairment? *Applied Psycholinguistics*, *21*, 159–181.

Bub, M. (2007). Fox, A. V. (2006). TROG-D: Test zur Überprüfung des Grammatikverständnisses [Testrezension]. *L.O.G.O.S. Interdisziplinär*, *15*, 156.

Clahsen, H. (1986). *Die Profilanalyse. Ein linguistisches Verfahren für die Sprachdiagnose im Vorschulalter* (Logotherapia, Bd. 3). Berlin: Marhold.

Cummins, J. (2008). BICS and CALP: Rationale and Status of the Distinction. In B. V. Street & N. H. Hornberger (Eds.), *Literacy – Encyclopedia of Language and Education Volume 2* (2nd ed., pp. 71–83). New York, NY: Springer. doi: 10.1007/978-0-387-30424-3_36

Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion* (Bachelorstudium Psychologie, Bd. 20). Göttingen: Hogrefe.

Elben, C. E. & Lohaus, A. (2000). *MSVK. Marburger Sprachverständnistest für Kinder*. Göttingen: Hogrefe.

Feldhusen, F., Brunner, M., Heinrich, C. & Pröschel, U. (2007). Anwendung des „Sprachverständnistests für komplexe syntaktische Strukturen (nach D. V. Bishop)“ bei 6- bis 8-jährigen Grundschülern. *HNO*, *55*, 729–736. doi: 10.1007/s00106-006-1531-3

Fox, A. V. (2013a). *TROG-D. Test zur Überprüfung des Grammatikverständnisses* (6. Aufl.; Erstaufl.: 2006). Idstein: Schulz-Kirchner.

Fox, A. V. (2013b). *Ihre Frage: Retest-Reliabilität des TROG-D* [Persönliche Mitteilung via Email vom 03. 06. 2013].

Gathercole, S. E., Willis, C. S., Baddeley, A. D. & Emslie, H. (1994). The children’s test of nonword repetition: A test of phonological working memory. *Memory*, *2*, 103–127. doi: 10.1080/09658219408258940

Goldhammer, F. & Hartig, J. (2012). Interpretation von Testresultaten und Testeichung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (Springer-Lehrbuch, 2., aktualisierte und überarbeitete Aufl., S. 173–201). Berlin: Springer. doi: 10.1007/978-3-642-20072-4_8

Grimm, H., Aktas, M. & Frevert, S. (2010). *SETK 3-5. Sprachentwicklungstest für drei- bis fünfjährige Kinder* (2., überarbeitete Aufl.). Göttingen: Hogrefe.

Grimm, H. & Schöler, H. (1998). *HSET. Heidelberger Sprachentwicklungstest* (2., verbesserte Aufl.). Göttingen: Hogrefe.

Grosche, M. & Volpe, R. J. (2013). Response-to-intervention (RTI) as a model to facilitate inclusion for students with learning and behaviour problems. *European Journal of Special Needs Education, 28,* 254–269. doi: 10.1080/08856257.2013.768452

Gruber, N. (2012). *TSVK - Test zum Satzverstehen von Kindern* (PSYNDEX Tests Review: 9006453). Zugriff am 26. 05. 2013 unter http://www.zpid.de/retrieval/PSYNDEXTests.php?id=9006453

IQWiG (Hrsg.). (2009). *Früherkennungsuntersuchung auf umschriebene Entwicklungsstörungen des Sprechens und der Sprache. Abschlussbericht* (Version 1.0). IQWiG-Berichte: S06-01. Zugriff am 30. 05. 2013 unter https://www.iqwig.de/download/S06-01_Abschlussbericht_Frueherkennung_umschriebener_Stoerungen_des_Sprechens_und_der_Sprache.pdf

Jones, F. W., Long, K. & Finlay, W. M. L. (2006). Assessing the reading comprehension of adults with learning disabilities. *Journal of Intellectual Disability Research*, *50*, 410–418. doi: 10.1111/j.1365-2788.2006.00787.x

Kannengieser, S. (2009). *Sprachentwicklungsstörungen. Grundlagen, Diagnostik und Therapie*. München: Elsevier.

Kendall, M. G. (1962). *Rank correlation methods*. London: Griffin.

Kuder, G. F. & Richardson, M. W. (1937). The Theory of the Estimation of Test Reliability. *Psychometrika*, *2*, 151–160. doi: 10.1007/BF02288391

Law, J., Boyle, J. & Harris, F. (1998). *Screening for speech and language delay. A systematic review of the literature* (Health Technology Assessment, Vol. 2). Southhampton: The National Coordinating Centre for Health.

Lenhard, W. & Marx, P. (2010) *Auswertungshilfe zur Berechnung von Testkennwerten von Screeningverfahren* [Computer software]. Würzburg: Universität Würzburg. Verfügbar unter http://www.psychometrica.de/Testkennwerte.xls

Little, R. J. A. & Rubin, D. B. (1987). *Statistical analysis with missing data* (Wiley Series in Probability and Mathematical Statistics). New York, NY: Wiley.

Lohmeier, K. (2007). TROG-D. Test zur Überprüfung des Grammatikverständnisses. Annette V. Fox [Rezensionen]. *Forum Logopädie*, *21*, 70.

Mahlau, K. & Blumenthal, Y. (2014). Zur inklusiven Förderung von GrundschülerInnen mit erhöhten sprachlichen Risiken. Erste Ergebnisse im Rahmen des Rügener Inklusionsmodells (RIM). *Logos*, *22*, 84–95.

Mahlau, K., Diehl, K., Voß, S. & Hartke, B. (2011). Das Rügener Inklusionsmodell (RIM). Konzeption einer inklusiven Grundschule. *Zeitschrift für Heilpädagogik*, *62*, 464–472.

Marx, H. (1992). Methodische und inhaltliche Argumente für und wider eine frühe Identifikation und Prädiktion von Lese-Rechtschreibschwierigkeiten. *Diagnostica*, *38*, 249–268.

Marx, P. & Lenhard, W. (2011). Diagnostische Merkmale von Screening-Verfahren zur Früherkennung möglicher Probleme beim Schriftspracherwerb. In M. Hasselhorn & W. Schneider (Hrsg.), *Frühprognose schulischer Kompetenzen* (Tests und Trends, Bd. 9, S. 68–84). Göttingen: Hogrefe.

Moosbrugger, H. (2012). Klassische Testtheorie (KTT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (Springer-Lehrbuch, 2., aktualisierte und überarbeitete Aufl., S. 173–201). Berlin: Springer.

Myers, T. A. (2011a). Goodbye, listwise deletion: Presenting Hot Deck Imputation as an easy and effective tool for handling missing data. *Communication Methods and Measures*, *5*, 297–310. doi: 10.1080/19312458.2011.624490

Myers, T. A. (2011b) *HOTDECK* [Computer software]. Fairfax. Verfügbar unter http://www.afhayes.com/public/hotdeck.sps

Neugebauer, U. & Becker-Mrotzek, M. (2013). *Die Qualität von Sprachstandsverfahren im Elementarbereich. Eine Analyse und Bewertung.* Köln: Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache. Verfügbar unter http://www.mercator-institut-sprachfoerderung.de/fileadmin/user_upload/Institut_Sprachfoerderung/Mercator-Institut_Qualitaet_Sprachstandsverfahren_Web.pdf

Pospeschill, M. (2010). *Testtheorie, Testkonstruktion, Testevaluation*. München: Reinhardt.

Purpura, J. E. (2012). Assessment of Grammar. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics.* Oxford: Blackwell. doi: 10.1002/9781405198431.wbeal0045

Purpura, J. E. (2013). Assessing Grammar. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (Vol. 1, pp. 100–124). Hoboken, NJ: Wiley-Blackwell. doi: 10.1002/9781118411360.wbcla147

Ricken, G., Fritz, A. & Balzer, L. (2011). MARKO-D: Mathematik und Rechnen – Test zur Erfassung von Konzepten im Vorschulalter. In M. Hasselhorn & W. Schneider (Hrsg.), *Frühprognose schulischer Kompetenzen* (Tests und Trends, Bd. 9, S. 127–146). Göttingen: Hogrefe.

Ricken, G., Fritz-Stratmann, A. & Balzer, L. (2013). *MARKO-D. Mathematik- und Rechenkonzepte im Vorschulalter – Diagnose*. Göttingen: Hogrefe.

Ritterfeld, U., Lüke, T. & Eiermann, N. D. (2013). *Medienbasierte Sprachförderung im Grundschulalter*. Unveröffentlichter Projektbericht. Dortmund: Technische Universität Dortmund.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (Psychologie Lehrbuch, 2., vollständig überarbeitete und erweiterte Aufl.). Bern: Huber.

Sarimski, K. (2013). Untersuchungen zur Validität des TROG-D in der Diagnostik sprachauffälliger Vorschulkinder. *Frühförderung Interdisziplinär*, *32*, 43–46.

Schmitz, P. & Fox, A. V. (2007). Sprachverstehenstest im Deutschen unter besonderer Berücksichtigung des TROG-D. *Forum Logopädie*, *21*, 18–25.

Siegmüller, J. (2010). Störungen der Grammatik. In J. Siegmüller & H. Bartels (Hrsg.), *Sprache – Sprechen – Stimme – Schlucken* (Leitfaden, 2., durchgesehene Aufl., S. 89–103). München: Elsevier.

Siegmüller, J., Kauschke, C., von Minnen, S. & Bittner, D. (2011). *TSVK. Test zum Satzverstehen von Kindern. Eine profilorientierte Diagnostik der Syntax*. München: Elsevier.

Sijtsma, K. & van der Ark, L. A. (2003). Investigation and Treatment of Missing Item Scores in Test and Questionnaire Data. *Multivariate Behavioral Research*, *38*, 505–528. doi: 10.1207/s15327906mbr3804_4

Slavin, R. E. (2002). Evidence-Based Education Policies: Transforming Educational Practice and Research. *Educational Researcher*, *31*, 15–21. doi: 10.3102/0013189X031007015

Spreer, M. (2013). Erfassung sprachlicher Fähigkeiten in inklusiven schulischen Settings – Beobachtungsmaterialien und Diagnoseverfahren im Überblick. Sprachdiagnostische Ziele: Erfassung sprachlicher Fähigkeiten auf unterschiedlichen Komplexitätsstufen Altersstufe: Schulalter, Schwerpunkt Grundschule. *Praxis Sprache*, *58*, 241–246.

Tröster, H. (2009). *Früherkennung im Kindes- und Jugendalter. Strategien bei Entwicklungs-, Lern- und Verhaltensstörungen*. Göttingen: Hogrefe.

Vossler-Thies, E., Stevens, A., Engel, R. R. & Licha, C. (2013). Erfassung negativer Antwortverzerrungen mit der deutschen Fassung des „Personality Assessment Inventory", dem „Verhaltens- und Erlebensinventar". *Diagnostica*, *59*, 73–85.

Watermeyer, M., Höhle, B. & Kauschke, C. (2011). Ausagieren von Sätzen versus Satz-Bild-Zuordnung: Vergleich zweier Methoden zur Untersuchung des Sprachverständnisses anhand von semantisch reversiblen Sätzen mit Objektvoranstellung bei drei- und fünfjährigen Kindern. In S. Hanne, T. Fritzsche, S. Ott & A. Adelt (Hrsg.), *Lesen lernen. Diagnostik und Therapie bei Störungen des Leseerwerbs* (Spektrum Patholinguistik, Bd. 4, S. 237–247). Potsdam: Universitätsverlag Potsdam.

Winkler, B. (2007). Fox, Annette V.: TROG-D: Test zur Überprüfung des Grammatikverständnisses. *Mitsprache*, *38*, 88–89.

Zimmermann, J., Benecke, C., Hörz, S., Rentrop, M., Peham, D., Bock, A. et al. (2013). Validierung einer deutschsprachigen 16-Item-Version des Inventars der Persönlichkeitsorganisation (IPO-16). *Diagnostica*, *59*, 3–16.

ZPID (Hrsg.). (2007). *TROG-D - Test zur Überprüfung des Grammatikverständnisses,* Universität Trier. PSYNDEX Tests Review: 9004807. Zugriff am 04. 06. 2013 unter http://www.zpid.de/retrieval/PSYNDEXTests.php?id=9004807

Tabellen

Tabelle 1

Vergleich der Itemschwierigkeit und part-whole-korrigierten Trennschärfe des TROG-D (Einzeltest) in der Normierung und dieser Studie

	Normierung (*N* = 870)[1]		**diese Studie (*N* = 93)**	
Itemblock	Schwierigkeit	Trennschärfe	Schwierigkeit	Trennschärfe
A	.99	.15	-[2]	-
B	.88	.39	-[2]	-
C	.91	.31	-[2]	-
D	.96	.35	-[2]	-
E	.84	.56	.99	.06
F	.90	.44	-[2]	-
G	.87	.55	.97	.25
H	.42	.49	.84	.09
I	.67	.67	.90	.21
J	.72	.54	.87	.41
K	.59	.60	.81	.32
L	.69	.69	.88	.32
M	.64	.67	.88	.33
N	.52	.66	.72	.40
O	.52	.59	.66	.37
P	.59	.69	.88	.45
Q	.23	.42	.30	.37
R	.38	.62	.62	.34
S	.06	.24	.06	.22
T	.39	.60	.67	.45
U	.26	.50	.35	.38

Anmerkungen. 1) Daten der Normierungsstudie aus: Fox, A. V. (2013). *TROG-D. Test zur Überprüfung des Grammatikverständnisses* (6. Auflage). Idstein: Schulz-Kirchner. 2) Die Itemblöcke A, B, C, D und F des Einzeltests wurden in der vorliegenden Studie von allen TeilnehmerInnen korrekt gelöst. Sie liefern daher keine Information.

Tabelle 2

Verteilungsmaße, Itemschwierigkeit, part-whole-korrigierte Trennschärfe des Gruppentests und Passung der Itemblöcke zur Gesamtskala

Itemblock	*M* (*SD*)	95% CI	Schiefe	Schwierigkeit	Trenn-schärfe	α ohne Itemblock	Block gelöst *M* (*SD*)
B	3.40 (1.14)	[3.16, 3.63]	-1.86	.85	.40	.81	.73 (.45)
C	3.48 (1.00)	[3.28, 3.69]	-2.05	.87	.44	.81	.73 (.45)
D	3.74 (0.55)	[3.63, 3.86]	-2.06	.94	.41	.81	.80 (.41)
E	3.74 (0.55)	[3.63, 3.86]	-2.06	.94	.22	.82	.80 (.41)
F	3.68 (0.75)	[3.52, 3.83]	-2.65	.92	.52	.81	.81 (.40)
G	3.89 (0.40)	[3.81, 3.98]	-4.94	.97	.33	.82	.91 (.28)
H	3.02 (0.88)	[2.84, 3.20]	-0.72	.76	.38	.81	.33 (.47)
I	3.76 (0.65)	[3.63, 3.90]	-3.62	.94	.36	.81	.84 (.37)
J	3.70 (0.57)	[3.58, 3.82]	-1.75	.92	.42	.81	.75 (.43)
K	3.47 (0.79)	[3.31, 3.64]	-1.47	.87	.53	.81	.62 (.49)
L	3.08 (0.68)	[2.94, 3.22]	-0.52	.77	.26	.82	.25 (.43)
M	3.56 (0.71)	[3.41, 3.71]	-1.68	.89	.28	.82	.67 (.47)
N	2.70 (1.04)	[2.48, 2.91]	-0.43	.67	.28	.82	.25 (.43)
O	2.81 (1.10)	[2.58, 3.03]	-0.52	.70	.47	.81	.34 (.48)
P	3.25 (0.92)	[3.06, 3.44]	-1.47	.81	.52	.81	.47 (.50)
Q	1.35 (1.14)	[1.12, 1.59]	0.53	.34	.24	.82	.04 (.20)
R	2.88 (1.43)	[2.59, 3.18]	-1.05	.72	.61	.80	.51 (.50)
S	1.48 (1.09)	[1.26, 1.71]	0.63	.37	.33	.82	.06 (.25)
T	3.28 (0.86)	[3.10, 3.46]	-0.89	.82	.38	.81	.52 (.50)
U	2.58 (0.95)	[2.39, 2.78]	-0.39	.65	.58	.80	.16 (.37)

Anmerkung. Itemblock A dient im Gruppentest als Beispiel zur Erklärung des Aufgabenformates und wurde deshalb aus den Analysen ausgeschlossen. „Block gelöst“ entspricht der Auswertung im TROG-D-Einzeltest (1=alle Items korrekt, 0=nicht alle Items korrekt)

Tabelle 3

Güteindizes der potentiellen Cut-Off-Werte im Vergleich

Cut-Off-Wert			**Kriterium**									
	RW		**TW ≤ 40**		**TW > 40**							
Z	**Kl. 2**	**Kl. 3**	**FN**	**RP**	**FP**	**RN**	**SQ**	**TQ**	**SN**	**SP**	**PK**	**YI**	**RATZ**
1	-.90	50	61	3	9	**8**	**73**	**18.3**	**88.2**	75.0	**90.1**	**52.9**	.65	0.69
2	-.25	56	65	**1**	**11**	20	61	33.3	77.4	**91.7**	75.3	35.5	**.67**	**0.88**

Anmerkungen. Z = Cut-Off-Wert des transformierten Testrohwertes, RW = Cut-Off-Wert des Testrohwertes nach Klassenstufe, FN = Anzahl falsch-negativer Befunde, RP = Anzahl richtig-positiver Befunde, FP = Anzahl falsch-positiver Befunde, RN = Anzahl richtig-negativer Befunde, SQ = Selektionsquote (Anteil positiver Befunde), TQ = Trefferquote (Anteil korrekter Befunde), SN = Sensitivität, SP = Spezifität, PK = Positive Korrektheit, YI = Youden-Index, RATZ = Relativer Anstieg der Trefferquote gegenüber der Zufallstrefferquote. Der im Vergleich bessere Wert ist fett gedruckt.

Abbildungen

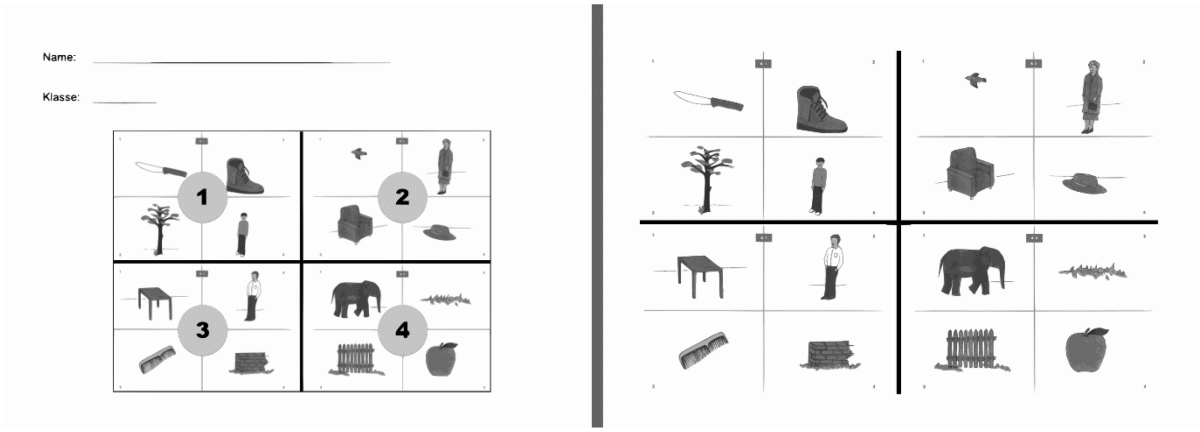

Abbildung 1. Deckblatt des Testheftes mit Übungsblock (links). Aufgabenseite aus dem Testheft des Gruppentests (rechts). Originalgrafiken mit freundlicher Genehmigung der Autorin und des Verlags: Fox (2013).

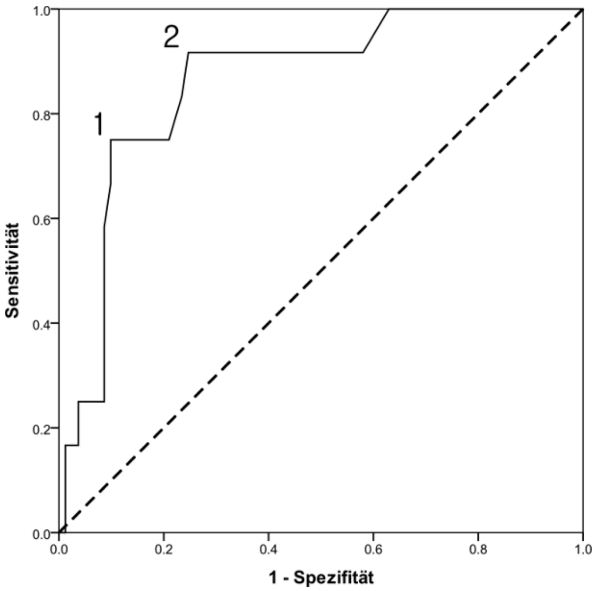

Abbildung 2. ROC-Kurve des Gruppenscreenings mit zwei potentiellen Cut-Off-Werten (1 & 2) und der AUC = .5-Diagonalen, die einer zufälligen Klassifikation entspricht.