Benchmarking

Deutsch B1-B2 PFLEGE und Deutsch B2-C1 MEDIZIN

24.–26.10.2013 in Köln

Beate Zeidler
Louise Lauppe
Testentwicklung/Qualitätsmanagement
l.lauppe@telc.net
Tel.: +49 (0) 69-956246-82

telc GmbH
Bleichstraße 1
60313 Frankfurt
www.telc.net

Deutsch B1-B2 PFLEGE

Die Prüfung telc Deutsch B1-B2 Pflege zielt darauf ab, Sprachkenntnisse von Pflegekräften, die in Deutschland praktizieren möchten, auf den Stufen des Gemeinsamen europäischen Referenzrahmens (GER) zu verorten. Es soll eine Aussage darüber getroffen werden, ob ein/e Prüfungsteilnehmer/in (TN) die Stufe B1 noch nicht erreicht hat, die Stufe B1 erreicht hat oder (mindestens) die Stufe B2 erreicht hat. Diese Feststellung wird für die rezeptiven Fertigkeiten, für das Schreiben und für das Sprechen getrennt getroffen.

Bei einer Veranstaltung vom 24. bis zum 26. Oktober 2013 wurde die Prüfung mit den Niveaustufen des GER in Relation gesetzt. Im Rahmen dieser Veranstaltung wurde unter anderem ein Benchmarking von TN-Performanzen im Sprechen durchgeführt, um abzugleichen, wie die TN anhand der Skalen der GER eingestuft werden und ob sie bei der Bewertung anhand der telc-Kriterien ein kompatibles Ergebnis erreichen. Die Konzeption der Veranstaltung erfolgte in Anlehnung an die vom Europarat empfohlenen Methoden (Council for Cultural Co-operation 2009). Im Folgenden wird beispielhaft die Bewertung eines TN (TN A) analysiert, um Schwierigkeiten bei der Bewertung aufzudecken und daraus Folgerungen für die Implementierung der Prüfung abzuleiten.

Der Prüfungsteil „Sprechen“

Dieser Prüfungsteil wird jeweils mit zwei TN durchgeführt und umfasst drei Abschnitte. In Teil 1 geht es um den Austausch von Meinungen und Erfahrungen. Jeder TN soll eine auf dem Aufgabenblatt abgebildete Situation kommentieren und wird anschließend durch den Prüfenden dazu befragt. Im Teil 2 präsentiert jeder TN eines von zwei Themen des Aufgabenblattes und beantwortet anschließend Rückfragen des Prüfenden. Im dritten Teil werden die TN aufgefordert, ihre Meinung zu dem Thema auf dem Aufgabenblatt auszudrücken und miteinander zu diskutieren.

Auswahl der Bewertergruppe

Die Bewertergruppe umfasste 22 Experten und Expertinnen, die den folgenden Personengruppen zugeordnet werden können: erfahrene und mit dem telc-System vertraute Prüfende, Unterrichtende im Bereich Deutsch für Pflegekräfte, Unterrichtende im Bereich Deutsch, Testkonstrukteure für Deutsch und/oder Deutsch Pflege.

Vorgehensweise und Ergebnisse

Die vom Europarat empfohlene Methodik sieht vor, zunächst eine Vertrautheit mit den Bewertungsskalen des GER herzustellen, dann einige bereits verortete Beispielvideos zu bewerten und danach, wenn eine hinreichende Konvergenz der Gruppe erreicht ist, die Videoaufnahmen der neuen, zu verortenden Prüfung bewerten zu lassen. In der Veranstaltung, über die berichtet wird, wurden zusätzlich durch dieselbe Bewertergruppe Bewertungen nach den telc-Kriterien vorgenommen, so dass ermittelt werden konnte, inwieweit die beiden Systeme zu gleichen Ergebnissen führen.

Zunächst wurde ein Video aus der Publikation „Mündlich“ (Bolton u.a. 2008) gezeigt und bewertet. „Mündlich“ enthält 21 Beispiele zu mündlicher Produktion und Interaktion für alle Niveaus des Gemeinsamen europäischen Referenzrahmens und wurde als Ergänzung zum GER in einem durch den Europarat geförderten Projekt erstellt. Gezeigt wurde die Interaktion Nr. 4 (Lisa und Christian)

und die Produktion Nr. 6 (Filippa). Die Leistungen der TN Christian und Filippa wurden anschließend folgendermaßen bewertet:

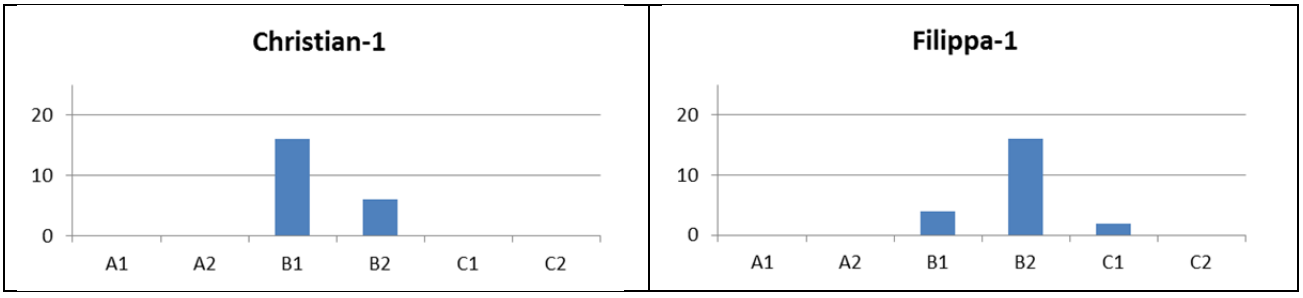
- In einer ersten Runde global nach den Deskriptoren des Gemeinsamen Europäischen Referenzrahmens (GER) „Mündliche Produktion/Interaktion allgemein“ (Council for Cultural Co-operation 2001, 64, 79)
- In einer zweiten Runde wiederum global nach GER-Deskriptoren „Mündliche Produktion/Interaktion allgemein“, nach Bekanntgabe der Ergebnisse der ersten Runde und Diskussion.

Dies diente der Vergegenwärtigung des GER-Konzepts der Stufen B1 und B2. TN Christian (Interaktion Nr. 4) repräsentiert in „Mündlich“ die Stufe B1 (Tendenz nach B1+ bei Flüssigkeit), TN Filippa (Produktion Nr. 6) die Stufe B2.

Zur Vorbereitung der Aktivität und um die Skalen den Bewerter/Innen bewusst zu machen, wurden zunächst die Skalen „Mündliche Produktion/Interaktion allgemein“ ausgegeben. Die Bewerter/Innen erhielten die Aufgabe, in der Skala „Mündliche Interaktion allgemein“ in Einzelarbeit die Schlüsselbegriffe zu unterstreichen. Die Ergebnisse wurden im Plenum zusammengetragen und besprochen.

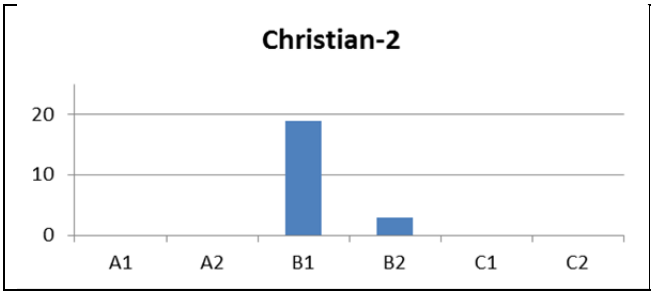
Danach wurden die Formblätter, auf denen die Bewertung abgegeben werden sollte, erläutert und das Video zum ersten Mal vorgespielt. Die Bewertungsblätter wurden eingesammelt und die Bewertung durch die Gruppe in folgender Form der Gruppe vorgestellt:

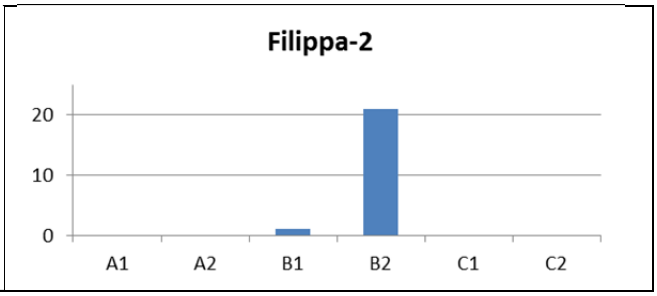
Anzahl Codes …	Christian-1	Filippa-1
A1	0	0
A2	0	0
B1	16	4
B2	6	16
C1	0	2
C2	0	0



Nach einer Diskussion der Stärken und Schwächen der beiden Performanzen erhielten die Bewerter/Innen die Gelegenheit, ihre Bewertung noch einmal zu revidieren. Das Ergebnis der zweiten Bewertungsrunde war:

Anzahl Codes	Christian-2	Filippa-2
A1	0	0
A2	0	0
B1	19	1
B2	3	21
C1	0	0
C2	0	0





Sowohl Christian als auch Filippa wurden in der zweiten Bewertungsrunde etwas strenger bewertet. Filippa erhält keine C1-Bewertungen mehr und Christian wird deutlicher der Stufe B1 zugeordnet.

Nach Abschluss dieser Runde und Bekanntgabe des Ergebnisses wurden die Kommentare zu den beiden Performanzen aus „Mündlich“ an die Bewertergruppe ausgegeben.

Die Bewertungen der Gruppe treffen recht gut die Niveaustufen, die die TN beispielhaft verdeutlichen sollen. Die Bedingung, maximal 1,5 Stufen Differenz bei der Einschätzung der Beispielperformanzen zu erreichen (Council for Cultural Co-operation 2009, 37), wurde erfüllt.

Nach weiterer Kalibrierung über TN-Performanzen aus „Mündlich“ wurden zwei Videos mit je zwei Teilnehmenden der Prüfung *telc Deutsch B1-B2 Pflege* von 20 Bewertern bewertet, und zwar:

- in der ersten Runde global nach GER-Deskriptoren „Mündliche Produktion/Interaktion allgemein“
- in der zweiten Runde global nach GER-Deskriptoren „Mündliche Produktion/Interaktion allgemein“, nach Bekanntgabe der Ergebnisse der ersten Runde und Diskussion
- in der dritten Runde nach telc-Kriterien
- in der vierten Runde, nach Bekanntgabe der Ergebnisse der dritten Runde und Diskussion, erhielten die Bewerter/Innen die Gelegenheit zur Revision ihrer Bewertung

Im Folgenden werden die Rohwerte aus den Bewertungen im Einzelnen dargestellt. Gezeigt werden die Durchschnittswerte für jedes Kriterium, sein Modus (=meistvergebener Eintrag), die jeweils höchste und niedrigste Bewertung, die Spannweite sowie die Standardabweichung. Die höchsten Werte sind rot, die niedrigsten blau dargestellt.

Die Bewertung nach dem telc-System umfasst Bewertungen für die folgenden Kriterien:

- Inhaltliche Angemessenheit Teil 1A (Beschreibung der abgebildeten Situation)
- Inhaltliche Angemessenheit Teil 1B (Austausch über Erfahrungen)
- Inhaltliche Angemessenheit Teil 2A (Präsentation)
- Inhaltliche Angemessenheit Teil 2B (Rückfragen)
- Inhaltliche Angemessenheit Teil 3 (Diskussion, Pausengespräch)
- Sprachliche Angemessenheit: Aussprache/Intonation
- Sprachliche Angemessenheit: Flüssigkeit
- Sprachliche Angemessenheit: Korrektheit
- Sprachliche Angemessenheit: Wortschatzbeherrschung

Für jedes Kriterium sind sechs Ausprägungen von „keine Leistung“ bis „gute B2-Leistung“ definiert. Die verwendeten Codes für die Ausprägungen sind: keine Leistung = 0, A2 = 1, B1 = 2, B1 gut = 3, B2 = 4, B2 gut = 5.

Unter den Bewertungen wird die Anzahl der Bewerter/innen dargestellt, die jeweils ein bestimmtes Urteil abgegeben haben, z.B. gaben in der dritten Runde bei TN A für den Inhalt in Teil 1B drei Bewerter/innen die Bewertung „B1 gut“, zwölf Bewerter/Innen die Bewertung „B2“ und fünf Bewerter/Innen die Bewertung „B2 gut“.

Diese Verteilung der Bewertungen ist darunter graphisch dargestellt.

Fehlende Daten sind als „.“ dargestellt.

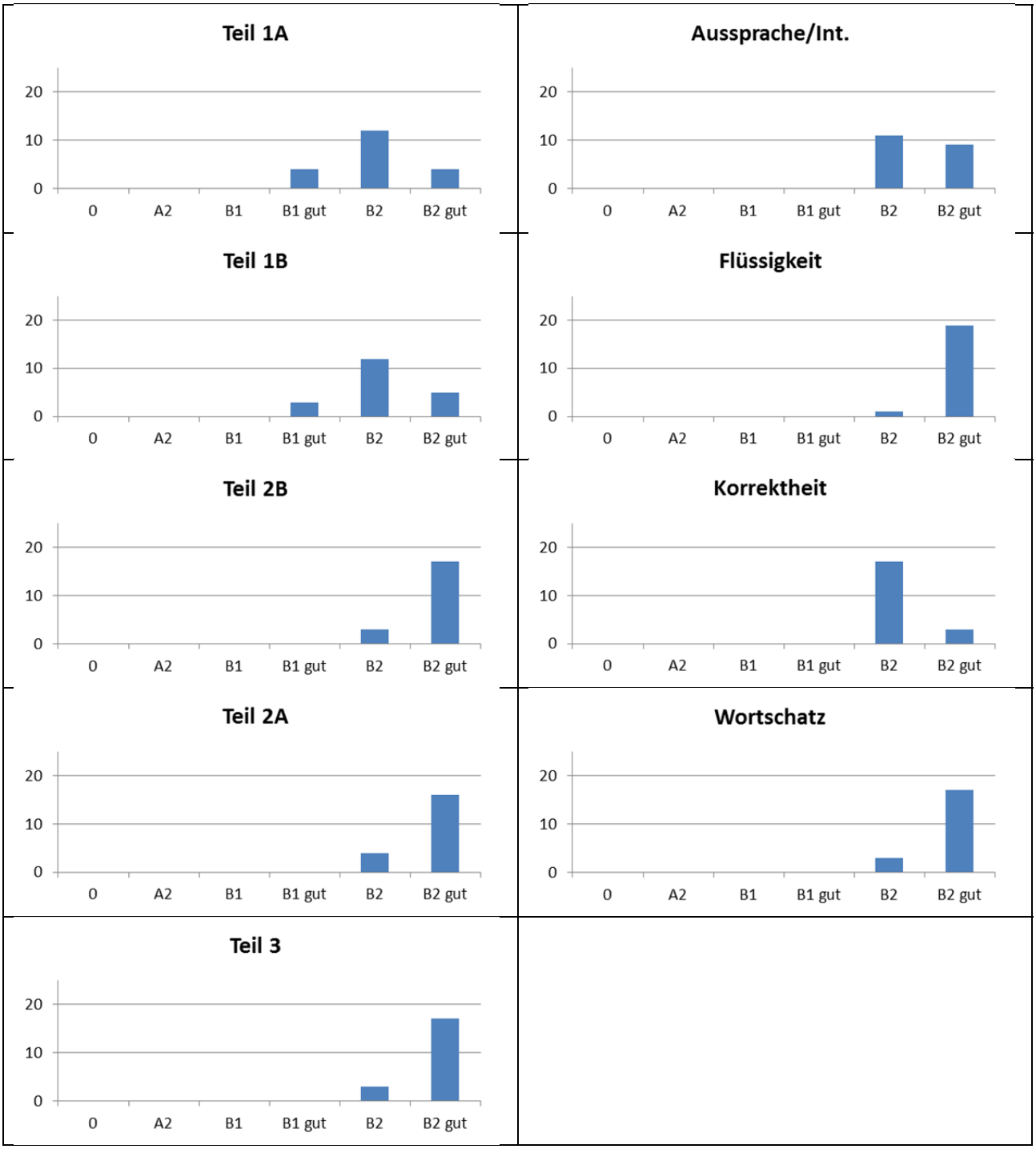
Globale Bewertungen nach GER-Kriterien

	erste Runde	*zweite Runde*
Anzahl Codes	**TN A**	**TN A**
A1	0	0
A2	0	0
B1	1	0
B2	19	19
C1	0	0
C2	0	0

Die zweite Bewertungsrunde bekräftigt eine eindeutige Verortung von TN A auf Stufe B2.

Bewertung nach telc-Kriterien, dritte Runde, TN A

Video 1, dritte Runde, TN A									
Durchschnitt	4	4,1	4,8	4,85	4,85	4,45	4,95	4,15	4,85
Modus	**4**	**4**	**5**	**5**	**5**	**4**	**5**	**4**	**5**
Max	5	5	5	5	5	5	5	5	5
Min	3	3	4	4	4	4	4	4	4
Spannweite	3	3	2	2	2	2	2	2	2
Standardabw	**0,63245553**	**0,6244998**	0,4	0,35707142	0,35707142	0,49749372	**0,21794495**	0,35707142	0,35707142
Anzahl Codes	**Teil 1A**	**Teil 1B**	**Teil 2A**	**Teil 2B**	**Teil 3**	**Aussprache/**	**Flüssigkeit**	**Korrektheit**	**Wortschatz**
0	0	0	0	0	0	0	0	0	0
A2	0	0	0	0	0	0	0	0	0
B1	0	0	0	0	0	0	0	0	0
B1 gut	4	3	0	0	0	0	0	0	0
B2	12	12	4	3	3	11	1	17	3
B2 gut	4	5	16	17	17	9	19	3	17



Die Bewertungen sind insgesamt sehr einheitlich, i.d.R. lässt sich eine eindeutige Zuordnung zu einer Stufe ableiten. Ausschließlich für den Inhalt von Teil 1 streuen die Bewertungen etwas breiter und bei Kriterium „Aussprache/Intonation“ erfolgt keine eindeutige Zuordnung des TN A zu einer bestimmten Stufe. Die Streuung in den Bewertungen bei Teil 1 lässt sich dadurch erklären, dass TN A verspätet zur Prüfung erschien, überstürzt mit der Prüfung beginnen musste und sich erst im Verlauf des ersten Prüfungsteils auf sein eigentliches Sprachniveau hocharbeitete.

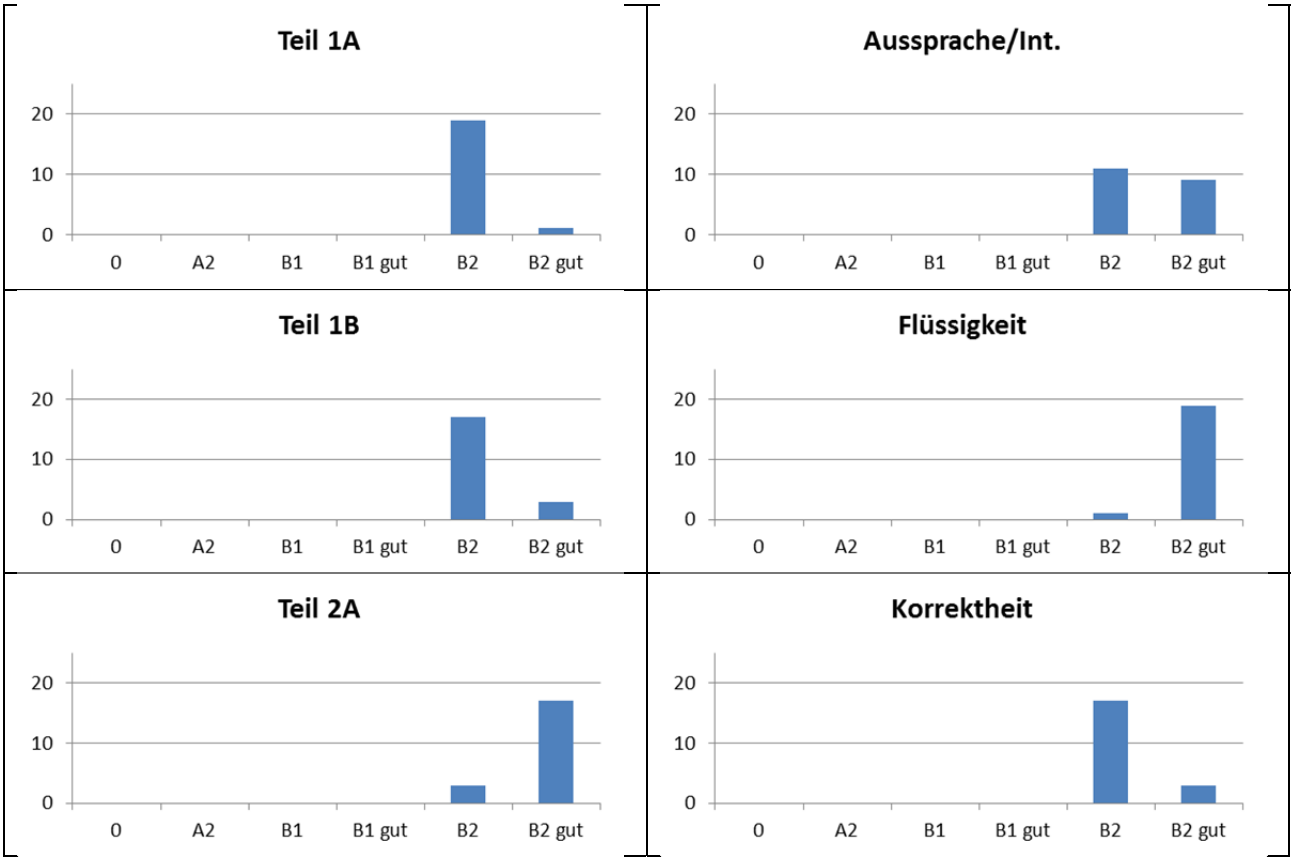
In der Diskussion standen zwei Themen im Vordergrund:

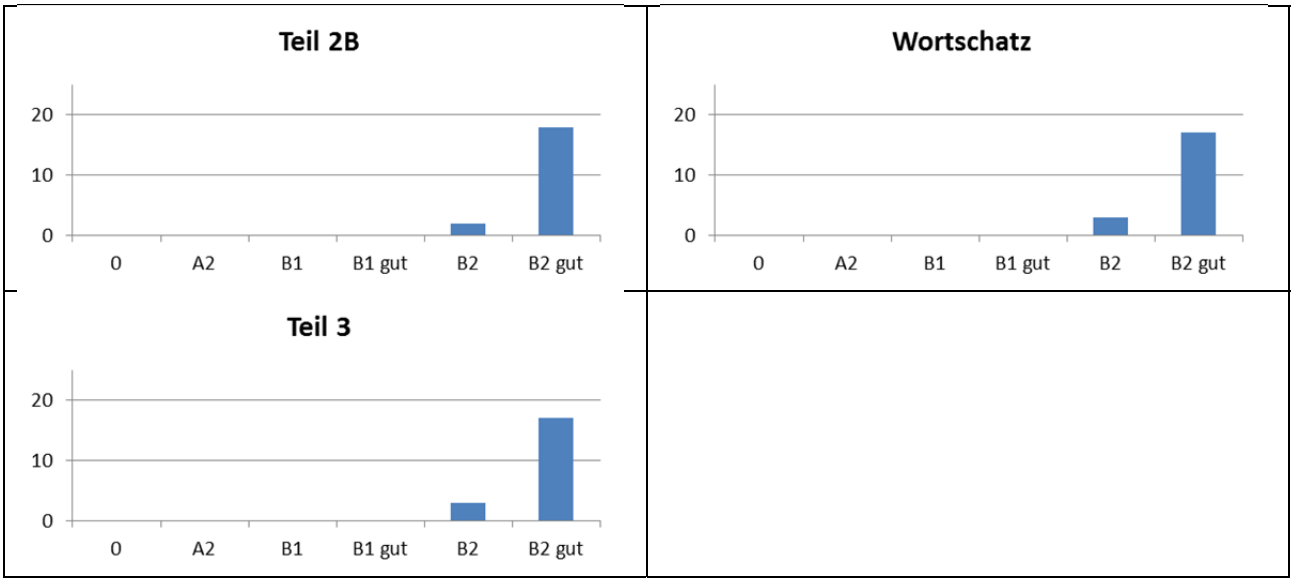
Das Kriterium „Wortschatzbeherrschung“ bezieht sich in allen Niveaustufen auf einen sogenannten Grundwortschatz. Neben dem allgemeinen Grundwortschatz soll in der Prüfung Deutsch Pflege jedoch auch der medizinisch-pflegerische Fachwortschatz und der entsprechende medizinisch-pflegerische Laienwortschatz geprüft werden, was in den Bewertungskriterien zu explizieren ist.

Da im Berufsalltag von Pflegekräften Missverständnisse fatale Folgen haben können, spielen Phonetik und Intonation eine bedeutende Rolle. Entsprechend wurde eine hohe Gewichtung des Kriteriums „Aussprache/Intonation“, insbesondere im Vergleich zu „Korrektheit“ für nötig erachtet.

Bewertung nach telc-Kriterien, vierte Runde, TN A

Video 1, vierte Runde, TN A									
Durchschnitt	4,05	4,15	4,85	4,9	4,85	4,45	4,95	4,15	4,85
Modus	**4**	**4**	**5**	**5**	**5**	**4**	**5**	**4**	**5**
Max	5	5	5	5	5	5	5	5	5
Min	4	4	4	4	4	4	4	4	4
Spannweite	2	2	2	2	2	2	2	2	2
Standardabw	0,21794495	**0,35707142**	**0,35707142**	0,3	**0,35707142**	**0,49749372**	**0,21794495**	**0,35707142**	**0,35707142**
Anzahl Codes	**Teil 1A**	**Teil 1B**	**Teil 2A**	**Teil 2B**	**Teil 3**	**Aussprache/**	**Flüssigkeit**	**Korrektheit**	**Wortschatz**
0	0	0	0	0	0	0	0	0	0
A2	0	0	0	0	0	0	0	0	0
B1	0	0	0	0	0	0	0	0	0
B1 gut	0	0	0	0	0	0	0	0	0
B2	19	17	3	2	3	11	1	17	3
B2 gut	1	3	17	18	17	9	19	3	17





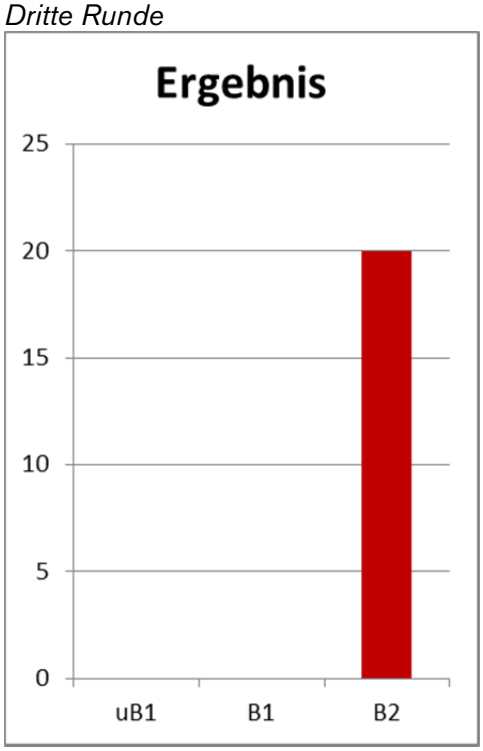
Video 1, dritte und vierte Runde, TN A: Gesamtergebnis

Bisher wurden die Bewertungen nach Kriterien getrennt betrachtet. Das TN-Ergebnis wird jedoch aus einer Zusammenschau aller neun Kriterien ermittelt. Dazu erhalten die Kriterien jeweils eine bestimmte Gewichtung. Diese ergibt sich aus dem Konstrukt der Prüfung und wird durch Experten/innen entsprechend den Anforderungen an die TN, deren Einlösung gemessen werden soll, festgelegt. Diese Gewichtung festzulegen, ist in jedem Fall Aufgabe der Prüfungseinrichtung, denn der GER enthält keine Vorgaben hinsichtlich der Gewichtung der Skalen.

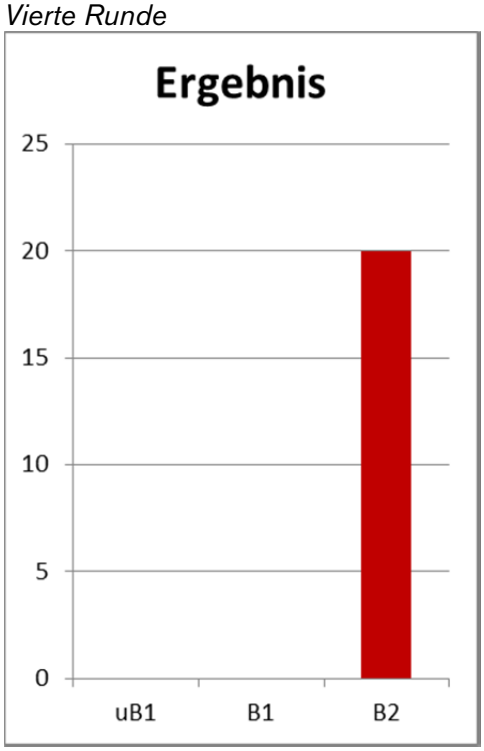
Nach Multiplikation mit den jeweiligen Gewichten werden die Punktzahlen addiert. Die Grenzwerte wurden aufgrund folgender Überlegung festgelegt: wenn ein TN in sieben von neun Kriterien eine Stufe erreicht (beispielsweise Stufe B2), so kann er/sie als B2-TN betrachtet werden (ausgenommen der Fall, dass die nicht bestandenen Kriterien das höchst gewichtete Kriterium (Faktor 4) bzw. zwei dreifach gewichtete Kriterien umfassen). Wird jedoch in drei Kriterien die Stufe um eine Stufe verfehlt (wobei die verfehlten Kriterien höchstens ein 1-fach und höchstens zwei 2-fach gewichtete Kriterien umfassen dürfen), so kann sie insgesamt nicht als erreicht betrachtet werden.

Das Gesamtergebnis der dritten und vierten Bewertungsrunde ist im Folgenden graphisch veranschaulicht.

Dritte Runde

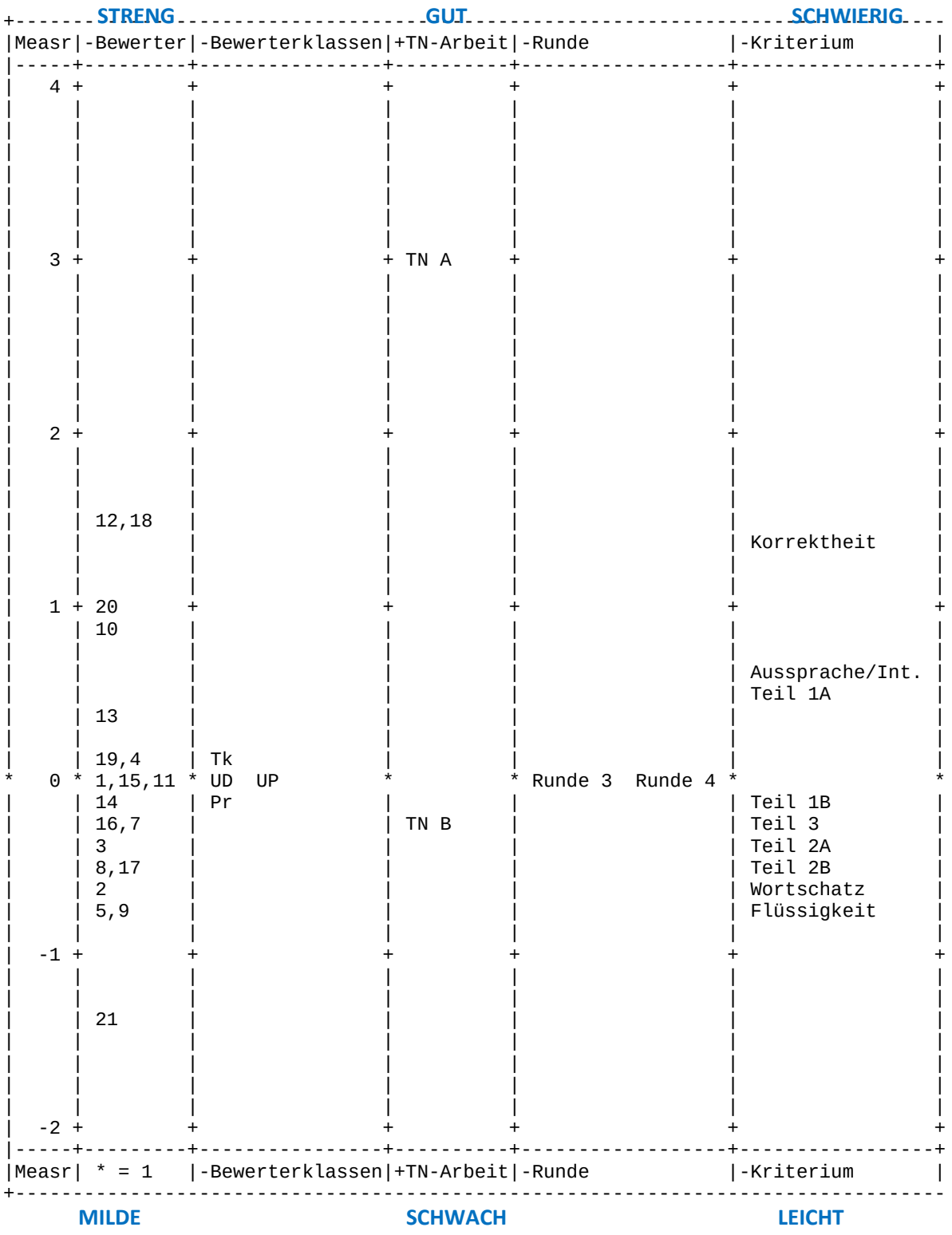


Vierte Runde



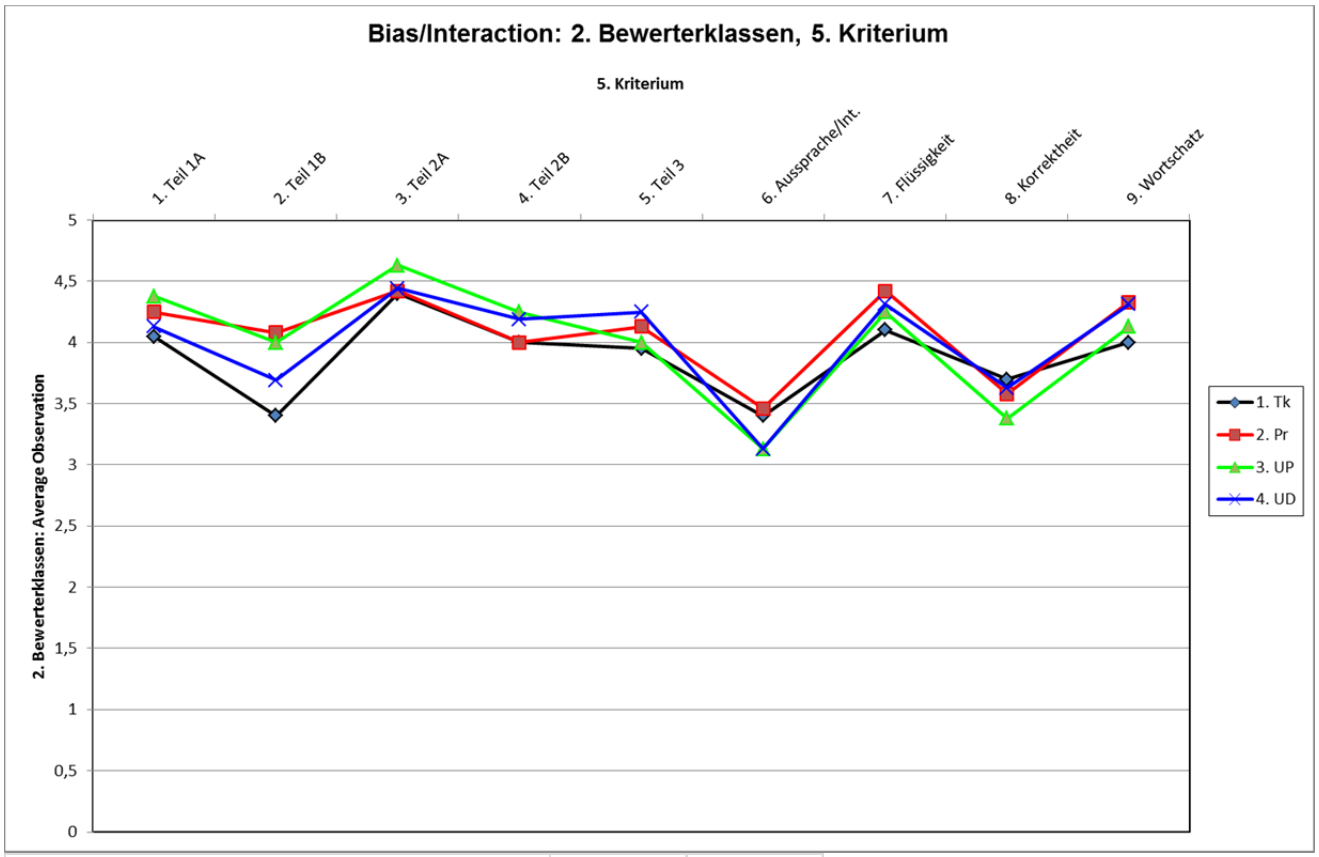
Trotz geringfügiger Veränderung der Bewertungen in den einzelnen Kriterien zwischen dritter und vierter Bewertungsrunde resultiert in beiden Bewertungsrunden eine einstimmige Verortung von TN A auf Niveaustufe B2.

Die Bewertungen wurden mit Hilfe der Software „Facets“ weitergehend analysiert. Die folgende Übersicht zeigt graphisch ein Gesamtbild der Bewertung von Video 1 – Strenge der Bewerter/Innen, Fähigkeit der Teilnehmenden, Schwierigkeit der Aufgaben und der Kriterien. Maßstab ist jeweils die Menge der Punkte, die vergeben wurden: ein strenger Bewerter vergibt wenige Punkte, ein fähiger TN erhält viele Punkte, auf eine schwierige Aufgabe und ein schwieriges Kriterium werden insgesamt wenige Punkte vergeben.



Ebenfalls in die Auswertung einbezogen wurden die Bewerterklassen sowie die Runde (dritte oder vierte Runde der Bewertung). Es wird deutlich, dass die Bewerter/Innen eine gewisse Spannweite hinsichtlich ihrer Strenge aufweisen. Die Bewerterklassen (Testkonstrukteure, Prüfende, Unterrichtende für Deutsch bzw. Deutsch Pflege) unterscheiden sich in dieser Hinsicht jedoch nicht wesentlich, ebenso zeigen sich keine wesentlichen Unterschiede in den Bewertungen zwischen Runde 3 und Runde 4. Die Kriterien sind jedoch deutlich unterschiedlich schwierig. Insbesondere bei „Korrektheit“ werden wenige Punkte erreicht, etwas schwieriger erscheint auch das Kriterium „Aussprache/Intonation“ sowie der Inhalt bei Teil 1A.

Gibt es Unterschiede in der Bewertung der Kriterien, die mit der Bewerterklasse zusammenhängen? Das folgende Diagramm zeigt die durchschnittlich durch die vier Bewerterklassen vergebenen, noch ungewichteten Punktwerte (es ist legitim, hier ungewichtete Punktwerte zu betrachten, da nicht das Gesamtergebnis, sondern gerade die Bewertung pro Kriterium im Fokus stehen). Beispielsweise vergab die Bewerterklasse Tk (Testkonstrukteure) in der Bewertung von Teil 1B durchschnittlich 3,4 ungewichtete Punkte und war somit etwas strenger als die anderen Bewerterklassen (UD: 3,69; UP: 4,0; Pr: 4,08).



5. Kriterium: t-value relative-to-overall (-)					
5. Kriterium		1. Tk	2. Pr	3. UP	4. UD
Teil 1A	1. Teil 1A	0,68	-0,12	-1,56	0,57
Teil 1B	2. Teil 1B	2,2	-1,36	-1,04	0,75
Teil 2A	3. Teil 2A	-0,28	0,8	-1,55	0
Teil 2B	4. Teil 2B	0,09	1,62	-0,88	-0,43
Teil 3	5. Teil 3	0,24	0,3	0,71	-1
Aussprache/	6. Aussprach	-1,58	-0,36	1,21	1,29
Flüssigkeit	7. Flüssigkeit	0,61	-0,58	0,28	-0,29
Korrektheit	8. Korrekthe	-1,98	0,31	1,34	-0,41
Wortschatz	9. Wortschat	0,13	-0,4	0,73	-0,76

Signifikante Unterschiede zeigten sich lediglich bei der Bewertung des Inhalts bei Teil 1B: Testkonstrukteure sind strenger in der Bewertung d.h. sie vergeben signifikant weniger Punkte, wobei die Unterschiede gering ausfallen (s.o.).

Abschließend soll noch die Modellanpassung der Kriterien betrachtet werden (inwiefern ermöglichen die Bewertungen pro Kriterium es, diesem Kriterium eine bestimmte Schwierigkeit mit einer gewissen Sicherheit zuzuordnen?).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	Kriterium
284	80	3.5	3.68	1.39	.20	.82	-1.1	.74	-1.5	1.19	.76	.71	8	Korrektheit
265	80	3.3	3.41	.65	.16	.75	-1.5	.81	-1.1	1.16	.87	.84	6	Aussprache/Int.
335	80	4.2	4.14	.55	.24	2.49	7.5	2.64	7.0	-1.00	-.07	.64	1	Teil 1A
306	80	3.8	3.95	-.10	.19	1.07	.4	1.18	1.0	.68	.54	.64	2	Teil 1B
327	80	4.1	4.26	-.30	.19	.61	-2.4	.53	-2.2	1.33	.87	.78	5	Teil 3
355	80	4.4	4.52	-.40	.24	.88	-.6	.85	-.5	1.11	.72	.68	3	Teil 2A
329	80	4.1	4.26	-.45	.20	.60	-2.5	.54	-2.4	1.36	.87	.77	4	Teil 2B
335	80	4.2	4.33	-.60	.20	.58	-2.8	.53	-2.8	1.44	.85	.73	9	Wortschatz
342	80	4.3	4.45	-.74	.20	.54	-3.0	.44	-3.0	1.48	.86	.73	7	Flüssigkeit
319.8	80.0	4.0	4.11	.00	.20	.93	-.7	.92	-.6		.70		Mean (Count: 9)	
27.5	.0	.3	.35	.67	.02	.58	3.1	.65	3.0		.29		S.D. (Population)	
29.2	.0	.4	.37	.71	.03	.61	3.3	.69	3.2		.31		S.D. (Sample)	

Model, Populn: RMSE .21 Adj (True) S.D. .64 Separation 3.11 Strata 4.48 Reliability .91
Model, Sample: RMSE .21 Adj (True) S.D. .68 Separation 3.32 Strata 4.75 Reliability .92
Model, Fixed (all same) chi-square: 102.1 d.f.: 8 significance (probability): .00
Model, Random (normal) chi-square: 7.4 d.f.: 7 significance (probability): .39

Die Bewertungen lassen sich durch die Modellannahmen zu Bewerterstrenge, TN-Fähigkeit und Kriterienschwierigkeit recht gut erklären. Eine Ausnahme bildet Kriterium „Inhaltliche Angemessenheit Teil 1A“. Wie oben dargelegt, spiegelt dies die Tatsache wider, dass TNA überstürzt mit der Prüfung beginnen musste, das tatsächliche Fähigkeitsniveau erst im Verlauf des ersten Prüfungsteils erreichte und daher kein eindeutiges Fähigkeitsniveau im Teil 1A zeigte.

Fazit

Die Bewertungen nach GER-Kriterien und nach telc-Kriterien insgesamt stimmen sehr gut überein. Die Gewichtung des Kriteriums „Aussprache/Intonation“ ist zu überarbeiten, um die zentrale Bedeutung dieser Facette herauszustellen und die TN-Leistung in diesem Bereich angemessen in das Prüfungsergebnis einfließen zu lassen. Das Kriterium „Wortschatzbeherrschung“ bedarf einer Ausdifferenzierung, was sich auf die Prüferqualifikation auswirkt und die Definition der Bewertungskriterien in den Prüferunterlagen betrifft.

Deutsch B2-C1 MEDIZIN

Die Prüfung telc Deutsch B2-C1 Medizin zielt darauf ab, Sprachkenntnisse von Medizinern, die in Deutschland praktizieren möchten, auf den Stufen des Gemeinsamen europäischen Referenzrahmens (GER) zu verorten. Es soll eine Aussage darüber getroffen werden, ob ein/e Prüfungsteilnehmer/in (TN) die Stufe B2 noch nicht erreicht hat, die Stufe B2 erreicht hat oder (mindestens) die Stufe C1 erreicht hat. Diese Feststellung wird für die rezeptiven Fertigkeiten, für das Schreiben und für das Sprechen getrennt getroffen.

Bei einer Veranstaltung vom 24. bis zum 26. Oktober 2013 wurde die Prüfung mit den Niveaustufen des GER in Relation gesetzt. Im Rahmen dieser Veranstaltung wurde ein Benchmarking von TN-Performanzen im Sprechen durchgeführt, um abzugleichen, wie die TN anhand der Skalen der GER eingestuft werden und ob sie bei der Bewertung anhand der telc-Kriterien ein kompatibles Ergebnis erreichen. Die Konzeption der Veranstaltung erfolgte in Anlehnung an die vom Europarat empfohlenen Methoden (Council for Cultural Co-operation 2009). Im Folgenden wird beispielhaft die Bewertung eines TN (TN A) analysiert, um Schwierigkeiten bei der Bewertung aufzudecken und daraus Folgerungen für die Implementierung der Prüfung abzuleiten.

Der Prüfungsteil „Sprechen“

Dieser Prüfungsteil wird jeweils mit zwei TN durchgeführt und umfasst drei Abschnitte. In Teil 1 geht es um ein Aufnahme- oder Anamnesegespräch. Einer der beiden Teilnehmenden nimmt die Rolle des Arztes ein, der andere die des Patienten, dann werden die Rollen getauscht. Im zweiten Teil stellt jede/r TN als Arzt den Patienten/die Patientin einem Kollegen/einer Kollegin vor und beantwortet Rückfragen dazu. Im dritten Teil spricht jeder der beiden TN über den in Teil 2 vorgestellten Patienten mit einem Angehörigen, der durch einen der beiden Prüfenden dargestellt wird (telc GmbH 2013).

Auswahl der Bewertergruppe

Die Bewertergruppe umfasste 22 Expert/Innen, die den folgenden Personengruppen zugeordnet werden können: erfahrene und mit dem telc-System vertraute Prüfende, Unterrichtende im Bereich Deutsch für Mediziner, Unterrichtende im Bereich Deutsch, Testkonstrukteure für Deutsch.

Vorbereitende Aktivitäten

Die vom Europarat empfohlene Methodik sieht vor, zunächst eine Vertrautheit mit den Bewertungsskalen des GER herzustellen, dann einige bereits verortete Beispielvideos zu bewerten und danach, wenn eine hinreichende Konvergenz der Gruppe erreicht ist, die Videoaufnahmen der neuen, zu verortenden Prüfung bewerten zu lassen. In der Veranstaltung, über die berichtet wird, wurden zusätzlich durch dieselbe Bewertergruppe Bewertungen nach den telc-Kriterien vorgenommen, so dass ermittelt werden konnte, inwieweit die beiden Systeme zu gleichen Ergebnissen führen.

Zunächst wurde ein Video aus der Publikation „Mündlich“ (Bolton u.a. 2008) gezeigt und bewertet. „Mündlich“ enthält 21 Beispiele zu mündlicher Produktion und Interaktion für alle Niveaus des Gemeinsamen europäischen Referenzrahmens und wurde als Ergänzung zum GER in einem durch den Europarat geförderten Projekt erstellt. Gezeigt wurden die TN Anne und Johanna (der dritte TN im Video, Aaron, wurde nicht bewertet), und zwar:

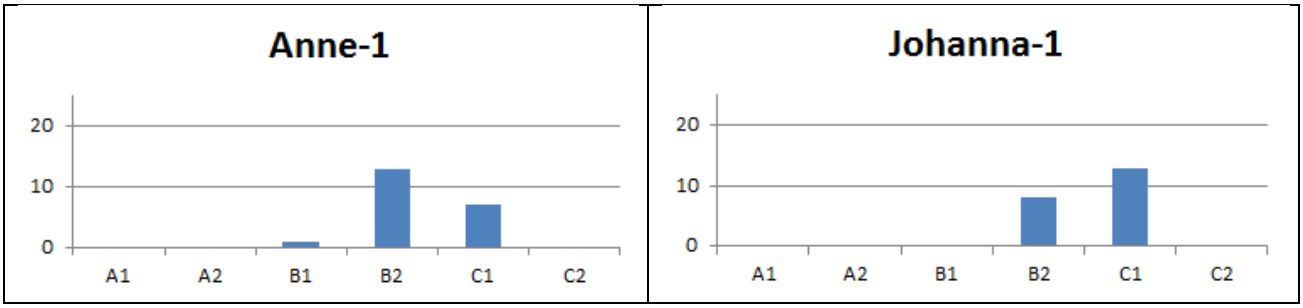
- in einer ersten Runde global nach den Deskriptoren des Gemeinsamen Europäischen Referenzrahmens (GER) „Mündliche Produktion/Interaktion allgemein“ (Council for Cultural Co-operation 2001, 64, 79)
- in einer zweiten Runde wiederum global nach GER-Deskriptoren „Mündliche Produktion/Interaktion allgemein“, nach Bekanntgabe der Ergebnisse der ersten Runde und Diskussion

Dies diente der Vergegenwärtigung des GER-Konzepts der Stufen B2 und C1. TN Anne repräsentiert in „Mündlich“ die Stufe B2+ (Tendenz nach C1 bei Flüssigkeit und Kohärenz), TN Johanna die Stufe C1 (Tendenz nach B2+ bei Korrektheit).

Zur Vorbereitung der Aktivität und um die Skalen den Bewertern/Innen bewusstzumachen, wurden zunächst die Skalen „Mündliche Produktion/Interaktion allgemein“ ausgegeben. Die Bewerter/Innen erhielten die Aufgabe, in der Skala „Mündliche Produktion allgemein“ in Einzelarbeit die Schlüsselbegriffe zu unterstreichen. Die Ergebnisse wurden im Plenum zusammengetragen und besprochen.

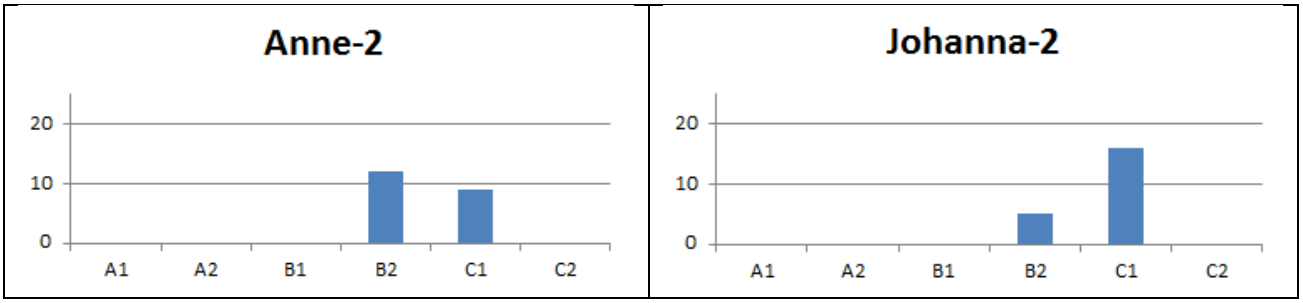
Danach wurden die Formblätter, auf denen die Bewertung abgegeben werden sollte, erläutert und das Video zum ersten Mal vorgespielt. Die Bewertungsblätter wurden eingesammelt und die Bewertung durch die Gruppe in folgender Form der Gruppe vorgestellt (eine Bewerterin konnte an dieser Runde nicht teilnehmen):

Anzahl Codes ...	Anne-1	Johanna-1
A1	0	0
A2	0	0
B1	1	0
B2	13	8
C1	7	13
C2	0	0



Nach einer Diskussion der Stärken und Schwächen der beiden Performanzen erhielten die Bewerter/Innen die Gelegenheit, ihre Bewertung noch einmal zu revidieren. Das Ergebnis dieser zweiten Bewertungsrunde war:

Anzahl Codes ...	Anne-2	Johanna-2
A1	0	0
A2	0	0
B1	0	0
B2	12	5
C1	9	16
C2	0	0



Sowohl Anne als auch Johanna wurden in der zweiten Bewertungsrunde besser bewertet. Anne erhält keine B1-Bewertung mehr, Johanna wird deutlicher der Stufe C1 zugeordnet.

Nach Abschluss dieser Runde und Bekanntgabe des Ergebnisses wurden die Kommentare zu den beiden Performanzen aus „Mündlich“ an die Bewertergruppe ausgegeben.

Die Bewertungen der Gruppe treffen recht gut die Niveaustufen, die die TN beispielhaft verdeutlichen sollen. Die Bedingung, maximal 1,5 Stufen Differenz bei der Einschätzung der Beispielperformanzen zu erreichen (Council for Cultural Co-operation 2009, 37), wurde erfüllt.

Nach weiterer Kalibrierung über TN-Performanzen aus „Mündlich“ wurde ein Video mit je zwei Teilnehmenden der Prüfung *telc Deutsch B2-C1 Medizin* von 22 Bewertern bewertet, und zwar

- in der ersten Runde global nach GER-Deskriptoren „Mündliche Produktion/Interaktion allgemein“
- in der zweiten Runde global nach GER-Deskriptoren „Mündliche Produktion/Interaktion allgemein“, nach Bekanntgabe der Ergebnisse der ersten Runde und Diskussion
- in der dritten Runde nach telc-Kriterien
- in der vierten Runde, nach Bekanntgabe der Ergebnisse der dritten Runde und Diskussion, erhielten die Bewerter/innen die Gelegenheit zur Revision ihrer Bewertung

Im Folgenden werden die Rohwerte aus den Bewertungen im Einzelnen dargestellt. Gezeigt werden die Durchschnittswerte für jedes Kriterium, sein Modus (=meistvergebener Eintrag), die jeweils höchste und niedrigste Bewertung, die Spannweite sowie die Standardabweichung. Die höchsten Werte sind rot, die niedrigsten blau dargestellt.

Die Bewertung nach dem telc-System umfasst Bewertungen für die folgenden Kriterien:

- Inhaltliche Angemessenheit Teil 1 (Patientengespräch)
- Inhaltliche Angemessenheit Teil 2A (Patientenvorstellung)
- Inhaltliche Angemessenheit Teil 2B (Rückfragen)
- Inhaltliche Angemessenheit Teil 3 (Gespräch mit Angehörigen)

- Sprachliche Angemessenheit: Aussprache/Intonation
- Sprachliche Angemessenheit: Flüssigkeit
- Sprachliche Angemessenheit: Korrektheit
- Sprachliche Angemessenheit: Wortschatzbeherrschung

Für jedes Kriterium sind sechs Ausprägungen von „keine Leistung“ bis „gute C1-Leistung“ definiert. Die verwendeten Codes für die Ausprägungen sind: keine Leistung = 0, B1 = 1, B2 = 2, B2 gut = 3, C1 = 4, C1 gut = 5.

Unter den Bewertungen wird die Anzahl der Bewerter/Innen dargestellt, die jeweils ein bestimmtes Urteil abgegeben haben, z.B. gab in der dritten Runde bei TN A für Teil 1, Kriterium 1, ein Bewerter die Bewertung „B1“, vier die Bewertung „B2“, usw.

Diese Verteilung der Bewertungen ist darunter graphisch dargestellt.

Fehlende Daten sind als „.“ dargestellt.

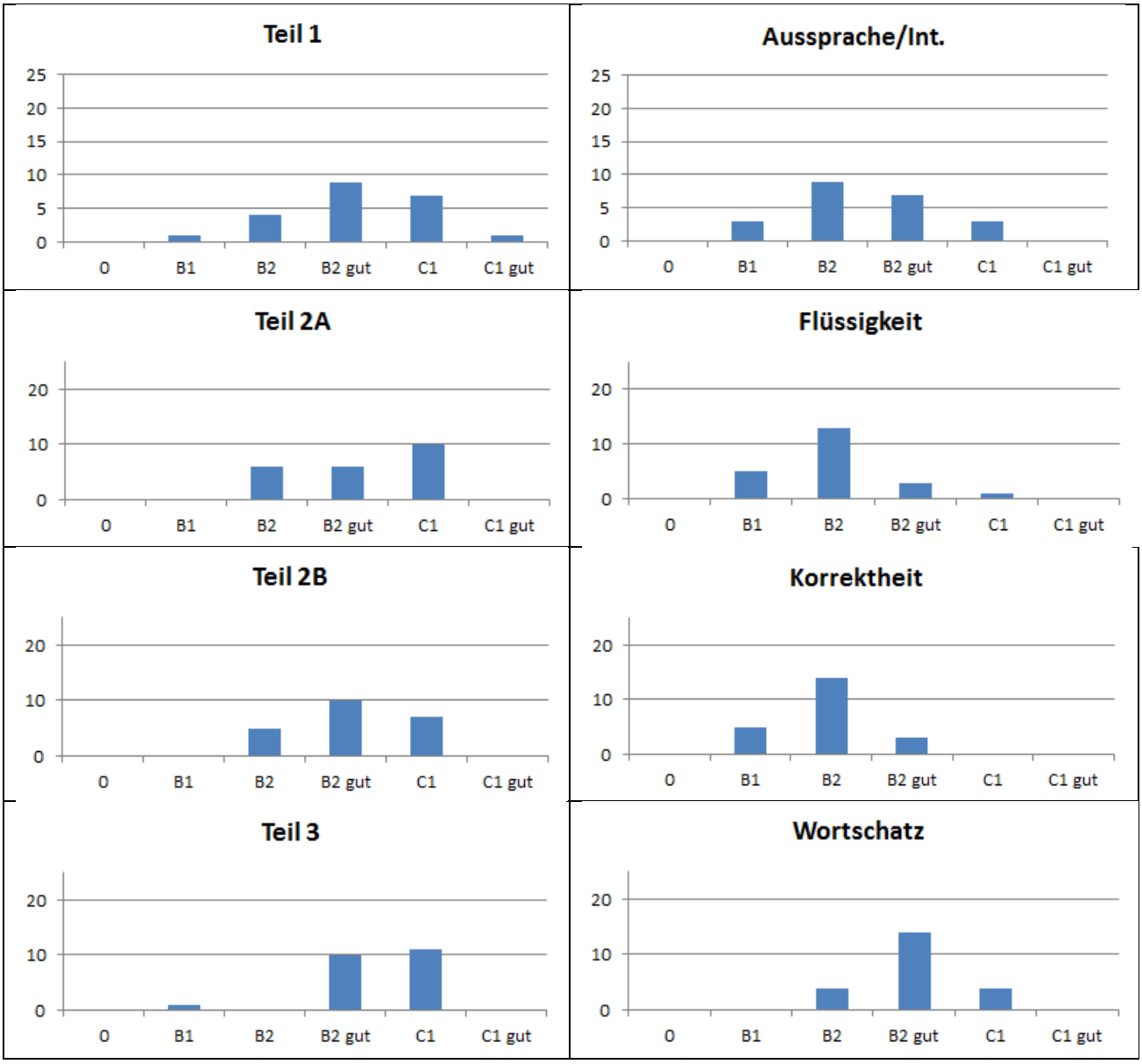
Globale Bewertung nach GER-Kriterien

	Erste Runde	*Zweite Runde*
Anzahl Codes ...	**TN A**	**TN A**
A1	0	0
A2	0	0
B1	6	1
B2	14	19
C1	2	2
C2	0	0

TN A wird relativ sicher auf Stufe B2 verortet, dies bekräftigt noch deutlicher die zweite Bewertungsrunde.

Bewertung nach telc-Kriterien, dritte Runde, TN A

Video 1, dritte Runde, TN A								
	Teil 1	**Teil 2A**	**Teil 2B**	**Teil 3**	**Aussprache/In**	**Flüssigkeit**	**Korrektheit**	**Wortschatz**
Durchschnitt	3,136363636	3,181818182	3,090909091	3,409090909	2,454545455	2	1,909090909	3
Modus	**3**	**4**	**3**	**4**	**2**	**2**	**2**	**3**
Max	5	4	4	4	4	4	3	4
Min	1	2	2	1	1	1	1	2
Spannweite	5	3	3	4	4	4	3	3
Standardabw.	0,919261292	0,833195581	0,732932523	0,717260629	0,890723543	0,738548946	0,596130775	0,603022689
Anzahl Codes ...	**Teil 1**	**Teil 2A**	**Teil 2B**	**Teil 3**	**Aussprache/In**	**Flüssigkeit**	**Korrektheit**	**Wortschatz**
0	0	0	0	0	0	0	0	0
B1	1	0	0	1	3	5	5	0
B2	4	6	5	0	9	13	14	4
B2 gut	9	6	10	10	7	3	3	14
C1	7	10	7	11	3	1	0	4
C1 gut	1	0	0	0	0	0	0	0



Besonders die Bewertungen für den Inhalt von Teil 1 und für Aussprache/Intonation streuen recht breit. Hohe Einigkeit herrscht dagegen bei den Kriterien „Korrektheit“, „Flüssigkeit“ und „Wortschatz“. Insgesamt scheint es schwieriger zu sein, die Bewertungen zu „Inhaltlicher Angemessenheit“ abzugeben als zu „Sprachlicher Angemessenheit“.

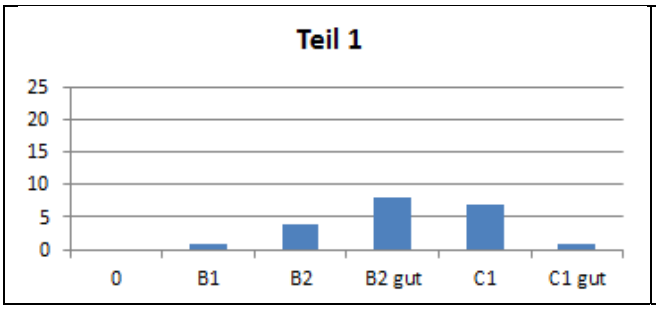
In der Diskussion traten folgende Fragestellungen hervor:

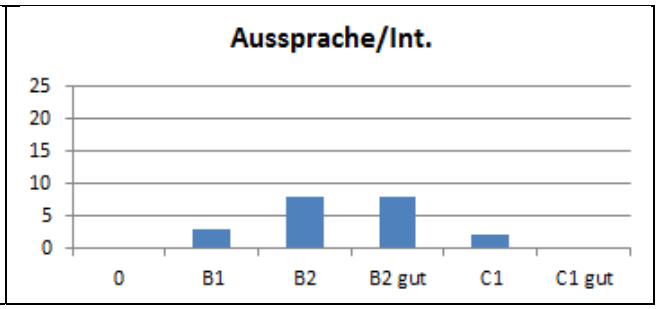
Inwieweit wird in Teil 1 die Patientenrolle in die Bewertung einbezogen? Vorgesehen ist, die TN-Performanz in dieser Rolle nur bei der Bewertung der „Sprachlichen Angemessenheit“ zu berücksichtigen, nicht aber bei der „Inhaltlichen Angemessenheit“. Dies stellt möglicherweise für ungeübte Prüfende eine besondere Schwierigkeit bei Teil 1 dar.

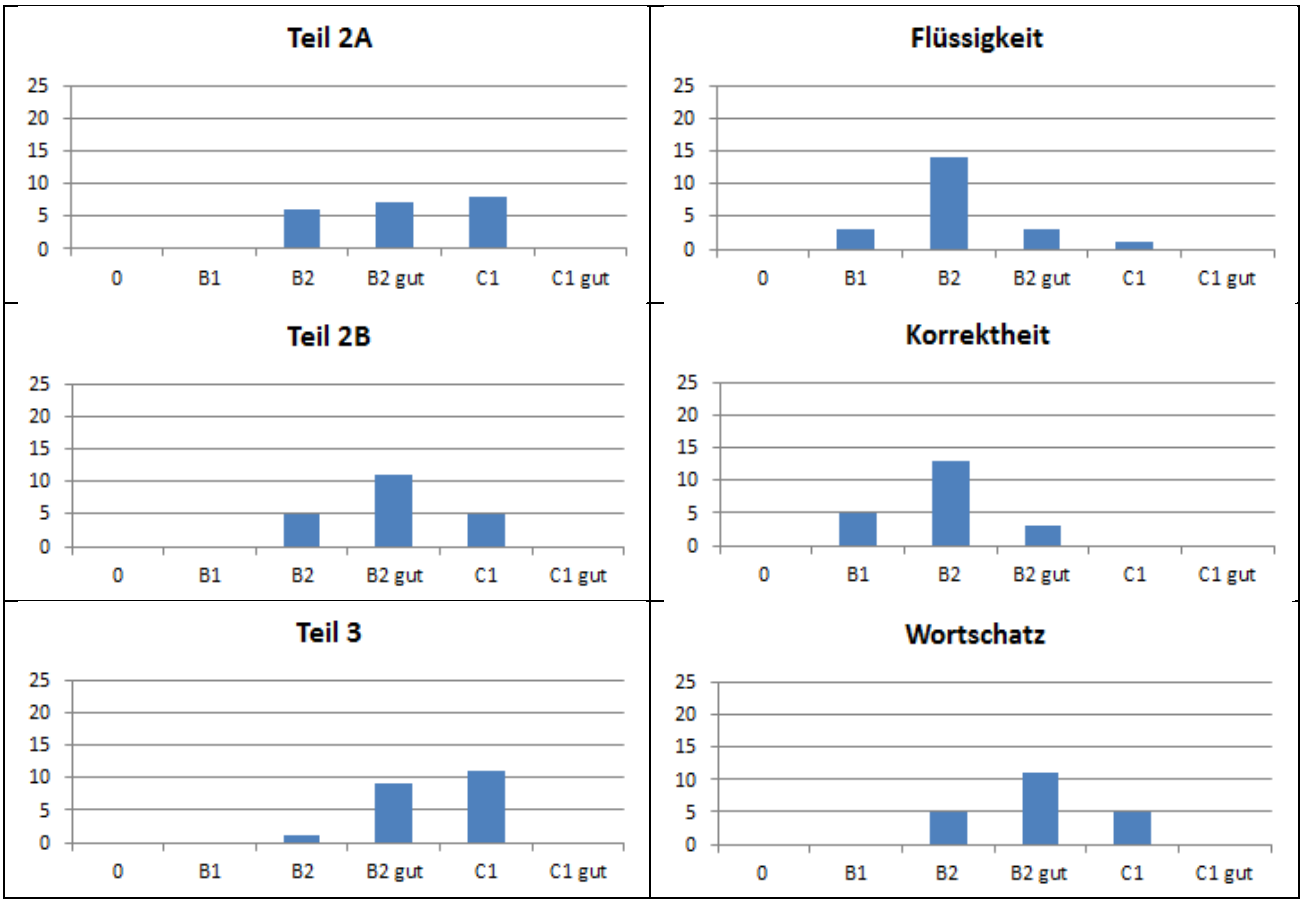
Ist die Verwendung von (möglicherweise auswendiggelernten) „Chunks“ zu sanktionieren? Es wurde klargestellt, dass Chunks, wenn sie *inhaltlich richtig* verwendet werden, keinen Grund zur Beanstandung darstellen.

Bewertung nach telc-Kriterien, vierte Runde, TN A

Video 1, vierte Runde, TN A								
	Teil 1	**Teil 2A**	**Teil 2B**	**Teil 3**	**Aussprache/I**	**Flüssigkeit**	**Korrektheit**	**Wortschatz**
Durchschnitt	3,14285714	3,0952381	3	3,47619048	2,42857143	2,0952381	1,9047619	3
Modus	**3**	**4**	**3**	**4**	**2**	**2**	**2**	**3**
Max	5	4	4	4	4	4	3	4
Min	1	2	2	2	1	1	1	2
Spannweite	5	3	3	3	4	4	3	3
Standardabw	0,94040084	0,81092316	0,69006556	0,58708705	0,84916926	0,68346191	0,60982136	0,69006556
inzahl Codes ..	**Teil 1**	**Teil 2A**	**Teil 2B**	**Teil 3**	**Aussprache/I**	**Flüssigkeit**	**Korrektheit**	**Wortschatz**
0	0	0	0	0	0	0	0	0
B1	1	0	0	0	3	3	5	0
B2	4	6	5	1	8	14	13	5
B2 gut	8	7	11	9	8	3	3	11
C1	7	8	5	11	2	1	0	5
C1 gut	1	0	0	0	0	0	0	0





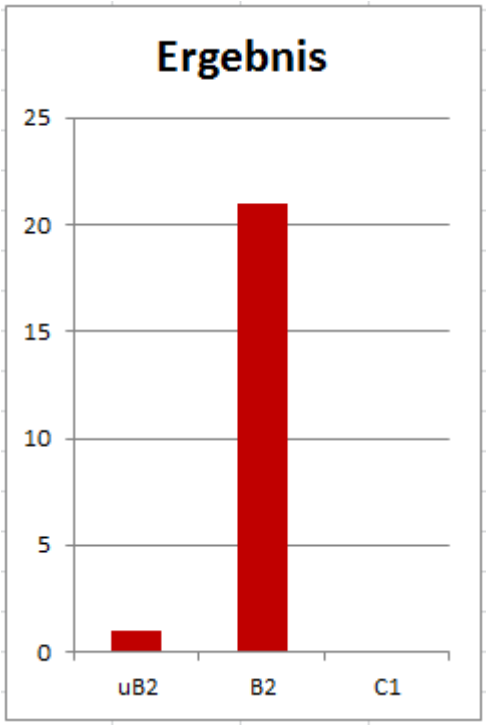


Video 1, Erste und zweite Runde, TN A: Gesamtergebnis

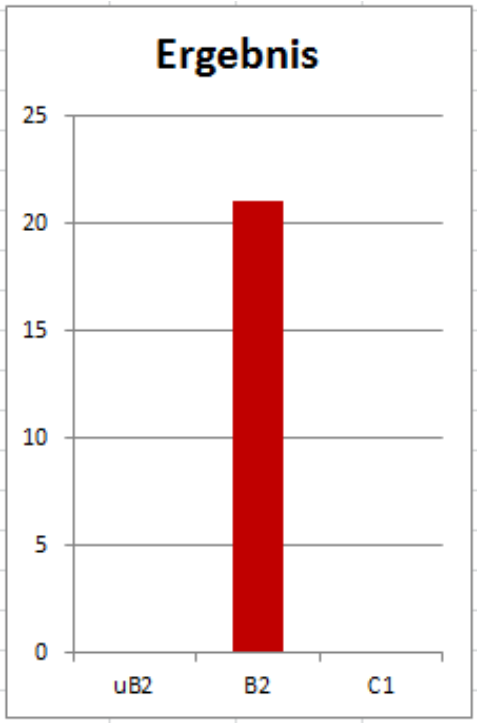
Bisher wurden die Bewertungen nach Kriterien getrennt betrachtet. Das TN-Ergebnis wird jedoch aus einer Zusammenschau aller acht Kriterien ermittelt. Dazu erhalten die Kriterien jeweils eine bestimmte Gewichtung. Diese ergibt sich aus dem Konstrukt der Prüfung und wird durch Experten entsprechend den Anforderungen an die TN, deren Einlösung gemessen werden soll, festgelegt. Diese Gewichtung festzulegen, ist in jedem Fall Aufgabe der Prüfungseinrichtung, denn der GER enthält keine Vorgaben hinsichtlich der Gewichtung der Skalen.

Nach Multiplikation mit den jeweiligen Gewichten werden die Punktzahlen addiert. Die Grenzwerte wurden aufgrund folgender Überlegung festgelegt: wenn ein TN in sechs von acht Kriterien eine Stufe erreicht (beispielsweise Stufe C1), so kann er/sie als C1-TN betrachtet werden. Wird jedoch in drei Kriterien die Stufe um eine Stufe verfehlt (oder in zwei Kriterien um mehr als eine Stufe), so kann sie insgesamt nicht als erreicht betrachtet werden. Bei den höchstgewichteten Kriterien (Inhalt Teil 3, Korrektheit, Wortschatz) kann jedoch nur maximal eines um eine Stufe verfehlt werden.

Erste Runde

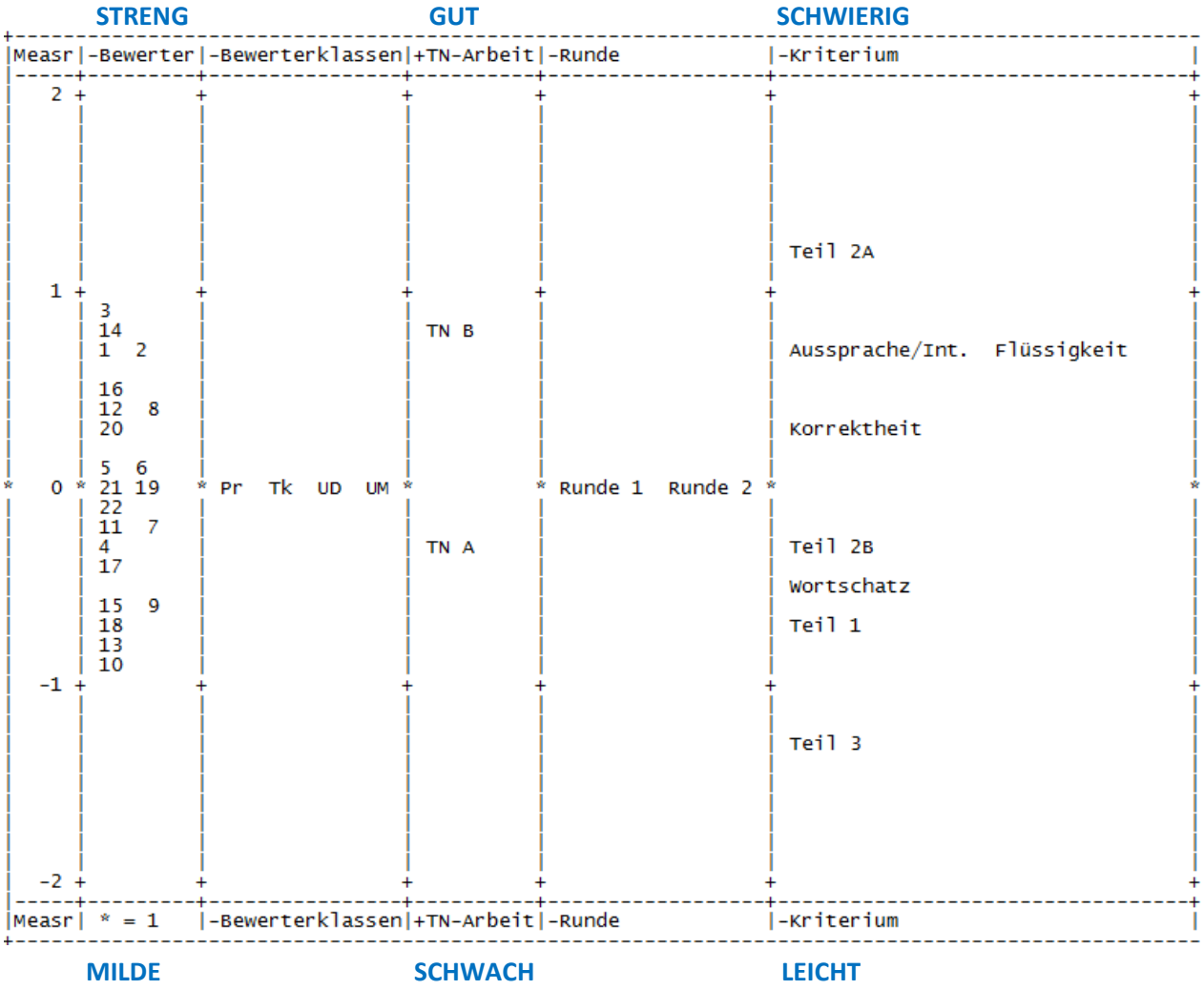


Zweite Runde



Während sich die Bewertung in den einzelnen Kriterien in der zweiten Runde nicht wesentlich verändert hat, wird TN A in der zweiten Runde einstimmig auf Niveau B2 verortet.

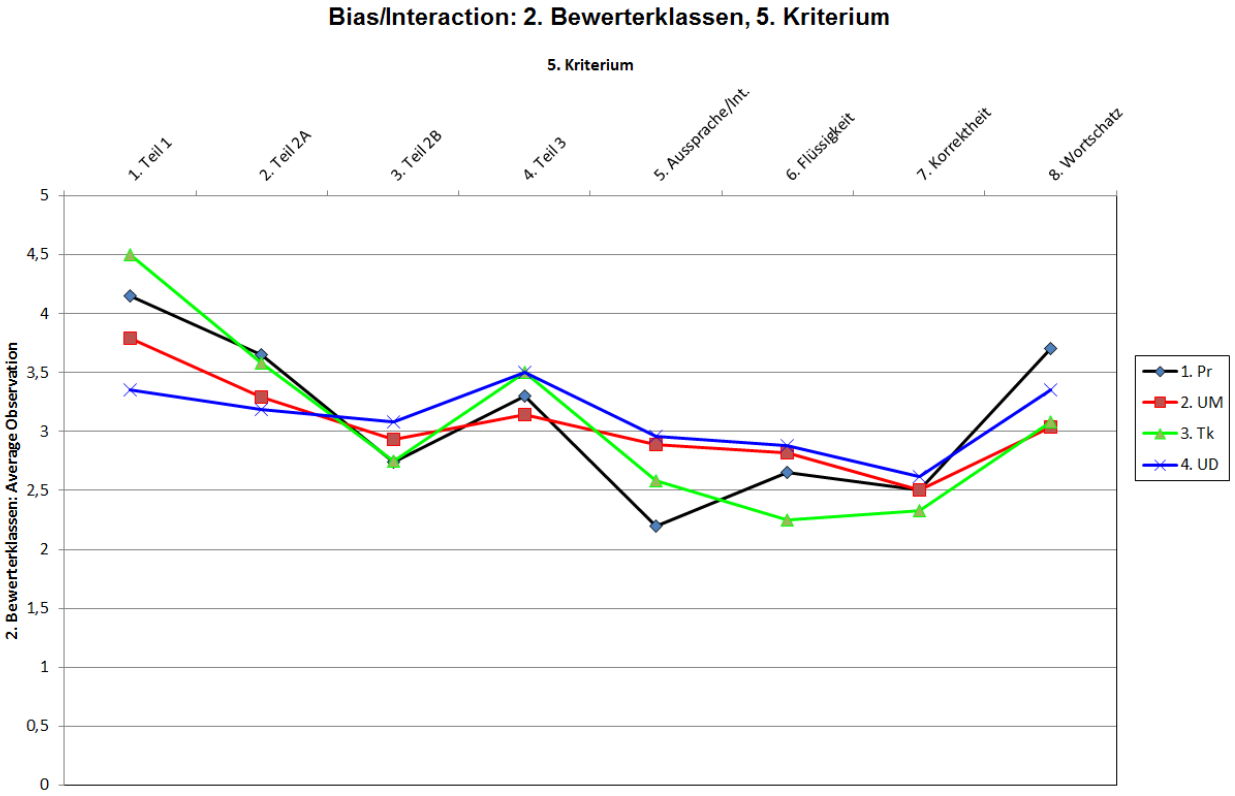
Die Bewertungen wurden mit Hilfe der Software „Facets“ weitergehend analysiert. Die folgende Übersicht zeigt graphisch ein Gesamtbild der Bewertung von Video 1 – Strenge der Bewerter/Innen, Fähigkeit der Teilnehmenden, Schwierigkeit der Aufgaben und der Kriterien. Maßstab ist jeweils die Menge der Punkte, die vergeben wurden: ein strenger Bewerter vergibt wenige Punkte, ein fähiger TN erhält viele Punkte, auf eine schwierige Aufgabe und ein schwieriges Kriterium werden insgesamt wenige Punkte vergeben.



Ebenfalls in die Auswertung einbezogen wurden die Bewerterklassen sowie die Runde (erste oder zweite Runde der Bewertung). Es wird deutlich, dass die Bewerter/Innen eine gewisse Spannweite hinsichtlich ihrer Strenge aufweisen. Die Bewerterklassen (erfahrene Prüfende, Testkonstrukteure, Unterrichtende für Deutsch bzw. Deutsch Medizin) unterscheiden sich in dieser Hinsicht jedoch nicht wesentlich, ebenso unterscheiden sich die Bewertungen in ihrer Strenge nicht nach Runde 1 oder 2. Die Kriterien sind jedoch deutlich unterschiedlich schwierig.

Gibt es Unterschiede in der Bewertung der Kriterien, die mit der Bewerterklasse zusammenhängen? Das folgende Diagramm zeigt die durchschnittlich durch die vier Bewerterklassen vergebenen, noch ungewichteten Punktwerte (es ist legitim, hier ungewichtete Punktwerte zu betrachten, da nicht das Gesamtergebnis, sondern gerade die Bewertungen pro Kriterium im Fokus stehen). Beispielsweise vergab die Bewerterklasse Tk (Testkonstrukteure) in der Bewertung von Teil 1 durchschnittlich 4,5 ungewichtete Punkte und war somit großzügiger als die Bewerterklasse Pr (Prüfende), die

durchschnittlich 4,15 Punkte vergab. Nicht bei jedem Kriterium sind jedoch signifikante Unterschiede festzustellen.



	5. Kriterium: t-value relative-to-overall (-)			
	1. Pr	2. UM	3. Tk	4. UD
1. Teil 1	-1,39	-0,05	-2,48	3,07
2. Teil 2A	-1,57	0,47	-1	1,61
3. Teil 2B	0,95	-0,41	0,68	-0,88
4. Teil 3	0,43	1,37	-0,91	-1,2
5. Aussprach	3,06	-1,5	0,43	-1,45
6. Flüssigkeit	0,55	-0,9	1,71	-0,75
7. Korrekthe	0,23	-0,16	0,74	-0,55
8. Wortschat	-2,8	2,01	1,12	-0,33

Signifikante Unterschiede zeigen sich wie folgt: Testkonstrukteure sind großzügiger in der Bewertung von Teil 1, Deutsch-Unterrichtende sind hier strenger. Im Kriterium „Aussprache“ sind Prüfende strenger als der Durchschnitt, dafür verhielten sie sich beim Wortschatz milder. Streng im Kriterium „Wortschatzbeherrschung“ waren insbesondere Unterrichtende aus dem Bereich Deutsch Medizin. An mögliche Interpretation kann vorgeschlagen werden:

Testkonstrukteure waren sich bereits darüber im Klaren, dass in Teil 1 die Darstellung der Patientenrolle keine Rolle spielt, und verhalten sich deshalb hier signifikant anders als die anderen Bewerter/Innen, die dies erst aus der Gruppendiskussion ableiten mussten.

Für Prüfende, die bisher nicht im Sonderbereich Medizin tätig waren, erscheint der Fachwortschatz, der in der Prüfung verwendet wird, wertvoller als für Unterrichtende aus diesem Bereich.

Daraus wären Folgerungen abzuleiten:

Die Bewertungsrelevanz der Patientenrolle für Teil 1 muss in den Prüferqualifizierungen deutlich herausgestellt werden.

Besondere Aufmerksamkeit in Prüferqualifizierungen verdient auch das Kriterium Wortschatz, da hier das (nicht bewertete) Fachwissen sehr eng mit der Sprachbeherrschung zusammenhängt.

Abschließend soll noch die Modellpassung der Kriterien betrachtet werden (inwiefern ermöglichen die Bewertungen pro Kriterium es, diesem Kriterium eine bestimmte Schwierigkeit mit einer gewissen Sicherheit zuzuordnen?)

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Kriterium
291	86	3.4	3.44	1.20	.16	.93	-.4	.87	-.7	1.07	.52	.48	2 Teil 2A
233	86	2.7	2.70	.72	.14	1.20	1.3	1.20	1.3	.77	.37	.52	5 Aussprache/Int.
234	86	2.7	2.67	.71	.13	.84	-1.1	.83	-1.2	1.18	.66	.56	6 Flüssigkeit
216	86	2.5	2.51	.28	.15	.79	-1.5	.79	-1.5	1.28	.65	.50	7 Korrektheit
247	85	2.9	2.97	-.31	.14	1.40	2.5	1.40	2.4	.47	.28	.52	3 Teil 2B
283	86	3.3	3.30	-.52	.18	.87	-1.0	.90	-.7	1.25	.56	.43	8 Wortschatz
330	86	3.8	3.91	-.75	.12	.78	-1.5	.83	-1.1	1.26	.68	.57	1 Teil 1
287	86	3.3	3.37	-1.35	.17	1.22	1.3	1.24	1.4	.73	.25	.44	4 Teil 3
265.1	85.9	3.1	3.11	.00	.15	1.00	-.1	1.01	.0		.50		Mean (Count: 8)
36.1	.3	.4	.45	.81	.02	.22	1.5	.22	1.4		.16		S.D. (Population)
38.6	.4	.4	.48	.87	.02	.24	1.6	.24	1.5		.17		S.D. (Sample)

Model, Popuin: RMSE .15 Adj (True) S.D. .80 Separation 5.34 Strata 7.45 Reliability .97
Model, Sample: RMSE .15 Adj (True) S.D. .86 Separation 5.72 Strata 7.96 Reliability .97
Model, Fixed (all same) chi-square: 229.8 d.f.: 7 significance (probability): .00
Model, Random (normal) chi-square: 6.8 d.f.: 6 significance (probability): .34

Die Bewertungen lassen sich durch die Modellannahmen zu Bewerterstrenge, TN-Fähigkeit und Kriterienschwierigkeit recht gut erklären. Eine Ausnahme bildet Kriterium „Inhaltliche Angemessenheit Teil 2B“. Hierzu wurde in der Diskussion erarbeitet, dass das Kriterium nur konsistent bewertbar ist, wenn die TN auch tatsächlich Rückfragen stellen. Das setzt voraus, dass ihnen diese Anforderung der Prüfung bewusst ist. Hieraus kann die Folgerung abgeleitet werden, die TN während der Prüfungsvorbereitung ausdrücklich darauf hinzuweisen.

Fazit

Die Bewertung nach GER-Kriterien und nach telc-Kriterien insgesamt stimmt überein. Für die Bewertung der Inhaltlichen Angemessenheit von Teil 1, Teil 2B und des Wortschatzes sollten weitere Maßnahmen getroffen werden, die sich hinsichtlich von Teil 1 und Wortschatz auf die Prüferqualifikation auswirken, hinsichtlich von Teil 2B auf die Gestaltung der TN-Unterlagen für die Prüfung.

Zitierte Publikationen

Bolton, Sibylle/ Glaboniat, Manuela/ Lorenz, Helga/ Müller, Martin, Perlmann-Balme, Michaela/ Steiner, Stefanie (2008): Mündlich. Mündliche Produktion und Interaktion Deutsch. Illustration der Niveaustufen des Gemeinsamen europäischen Referenzrahmens, Berlin/München/Wien/Zürich/New York (Langenscheidt)

Council for Cultural Co-operation, Education Committee, Modern Languages Division (2001): Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen, Berlin/München/Wien/Zürich/New York (Langenscheidt)

Council for Cultural Co-operation, Education Committee, Modern Languages Division (2009): Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual, Strasbourg (Council of Europe)

telc GmbH, telc Deutsch B2-C1 Medizin. Übungstest 1, Frankfurt (telc) 2013

telc GmbH, telc Deutsch B1-B2 Pflege. Übungstest 1, Frankfurt (telc) 2013.