

# Schnell zu erfassen: Ein Komplexitätsmaß für Mehrfachauswahlfragen

Frauke Kämmerer<sup>1</sup>, Heinrich Söbke<sup>2</sup> und Lukas Hartung<sup>2</sup>

**Abstract:** Mehrfachauswahlfragen können für Lern- und Prüfungszwecke eingesetzt werden. Bei der Nutzung zu Unterhaltungszwecken – wie in Quiz-Apps – ist eine um ein Vielfaches größere Popularität zu beobachten. Eine Analyse möglicher Ursachen für die hohe Attraktivität lässt u.a. eine geringere textuelle Komplexität der Fragen in Unterhaltungskontexten vermuten. Auf Grundlage existierender textueller Komplexitätsmaße haben wir daher eine Metrik für die textuelle Komplexität einer Mehrfachauswahlfrage entwickelt. Zur Validierung wurde die entstandene Metrik analytisch auf zwei Corpora von Mehrfachauswahlfragen angewendet. Die Fragen des Corpus mit Fachfragen zeigten sich deutlich komplexer als die Fragen des Unterhaltungs-Corpus. Darauf aufbauend wurden die Fachfragen so überarbeitet, dass sie entsprechend der Metrik eine geringere Komplexität zeigten. Alle drei Gruppen von Fragen wurden in einer Probanden-Studie (n=14) wettbewerbsorientiert in einer Quiz-App beantwortet. Obwohl eine Annäherung erzielt werden konnte, wiesen die Unterhaltungsfragen weiterhin bessere Werte bzgl. des Spielspaßes auf. Wesentliche Ergebnisse der Arbeit sind ein Entwurf einer Komplexitätsmetrik für Mehrfachauswahlfragen, deren Validierung und studiengestützte Hinweise, dass eine konstruktive Nutzung der Metrik zu leichter lesbaren Fragen führt.

**Keywords:** Multiple Choice Question; Metrik; Quiz; Instructional Design; Lesbarkeitsindex

## 1 Einleitung

Mehrfachauswahlfragen (Multiple-Choice-Fragen, MCQ) können z. B. zur Überprüfung des Lernfortschritts eingesetzt werden. Gleichzeitig lassen sie sich auch zum Lernen nutzen, denn durch wiederholtes Beantworten von Fragen werden Lernprozesse initiiert („Testing-effect“, [BR07]). Eine derartige Fragenwiederholung lässt sich auch in kommerziellen Quiz-Apps beobachten [Ha14], die z. T. hohe Benutzerzahlen aufweisen [Ru14]. In früheren Arbeiten haben wir beobachtet, dass Spieler von Quiz-Apps diese mit einer Lernerwartung und -bereitschaft nutzen [Sö15]. Gleichzeitig wird ein teilweise vorhandener Zeitdruck, beispielsweise ein Zeitlimit von zehn Sekunden in der App *QuizUp* [Pl14], bei der Beantwortung der Fragen als eher abträglich für den Lernerfolg gesehen, da Fachfragen tendenziell länger und insgesamt schwerer erfassbar zu sein scheinen als Fragen der Unterhaltungsthemen [SW16]. Tabelle 1 zeigt jeweils ein Beispiel einer Unterhaltungsfrage des Themengebietes *Allgemeinwissen* und eine Fachfrage des Themas *Technische Luftreinhaltung*. Es ist anzunehmen, dass die

---

<sup>1</sup> Hochschule Nordhausen, Weinberghof 4, 99734 Nordhausen, frauke.kaemmerer@fh-nordhausen.de

<sup>2</sup> Bauhaus-Universität Weimar, Bauhaus-Institut für zukunftsfähige Infrastruktursysteme, Coudraystr. 7, 99423 Weimar, {heinrich.soebke|lukas.hartung}@uni-weimar.de

unterschiedliche textuelle Komplexität nur eine mögliche Ursache für das Zurückbleiben der Akzeptanz von kommerziellen Quiz-Apps in Lernkontexten hinter deren Popularität als Unterhaltungsmedium ist. In multikausalen Erklärungsmodellen spielen sicherlich auch weitere Gründe wie unterschiedliche Nutzerpräferenzen, der Lernkontext selbst sowie die Beschränkung auf festgelegte Themen eine Rolle. Jedoch lässt sich für die *Textuelle Komplexität* mit Metriken in Form von objektiv feststellbaren Kennzahlen analytisch (bei existierenden Fragen) und konstruktiv (beim Fragenentwurf) automatisiert mit relativ geringem Aufwand Abhilfe schaffen [SK16].

| <b><u>Thema:</u></b> Allgemeinwissen         | <b><u>Thema:</u></b> Technische Luftreinhaltung                                               |
|----------------------------------------------|-----------------------------------------------------------------------------------------------|
| Aus welcher Bohnenart wird Tofu hergestellt? | Welche Geschwindigkeit wird verwendet, um den Widerstandsbeiwert bei Biofiltern zu berechnen? |
| <u>A: Sojabohnen</u>                         | <u>A: Anströmungsgeschwindigkeit</u>                                                          |
| B: Weiße Bohnen                              | B: Durchschnittsgeschwindigkeit                                                               |
| C: Grüne Bohnen                              | C: Ausströmungsgeschwindigkeit                                                                |
| D: Kidneybohnen                              | D: Schallgeschwindigkeit                                                                      |

Tabelle 1 Beispiel einer Unterhaltungs- und einer Fachfrage

## 2    Textuelle Komplexität und Komplexitätskriterien

### 2.1    Grundlagen

Metriken zur Messung des Leseaufwandes von Auswahlfragen scheinen bisher nicht im Fokus der Lehr-Lern-Forschung gestanden zu haben. So finden sich zwar Richtlinien für den Fragenentwurf (z.B. [HR13]), nicht aber konkrete Kennzahlen. Hingegen gibt es für Texte im Allgemeinen mehrere Lösungsansätze. Bamberger und Vanecek [BV84] beschreiben Lesbarkeitsformeln, die formale Textmerkmale in eine Maßzahl für die Verständlichkeit abbilden. Um zu einem ersten Entwurf einer Metrik zu gelangen, wird hier vereinfachend davon ausgegangen, dass Leseaufwand und Verständlichkeit miteinander korrelieren. Oft kann eine Metrik nicht ohne Anpassung auf eine andere Sprache übertragen werden, da Textmerkmale genutzt werden, die von Sprache zu Sprache divergieren.

In der Softwareentwicklung ist es gleichfalls essentiell, Quelltexte möglichst schnell und korrekt erfassen zu können. Da es sich hier um formal strukturierte Texte handelt, sind Metriken für diese besonders leicht zu entwerfen und anzuwenden.

Sowohl für die Lesbarkeit von Texten als auch für die Komplexität von Quellcode gibt es eine Vielzahl von Metriken. Daher haben wir uns im Rahmen dieser Arbeit auf fünf Verfahren beschränkt, die sich im Rahmen der Literaturrecherche als relevant herausstellten. Diese wurden bezüglich Eingangsgrößen, Rechenvorschriften und Ausgangswert analysiert (s. Tabelle 2).

| Metrik                           | Beschreibung                                                                                                                        | Wesentliche Eingangsgrößen                                                                                                                                                                                               |
|----------------------------------|-------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Flesch-Reading-Ease-Formel (FRE) | Maßzahl zur Berechnung der Lesbarkeit eines Textes [Fl48], deutsche Bearbeitung [Am78]                                              | <ul style="list-style-type: none"> <li>Durchschnittliche Satzlänge</li> <li>Durchschnittliche Silbenanzahl pro Wort</li> </ul>                                                                                           |
| Gunning-Fog-Index (GFI)          | Maßzahl zur Beschreibung der Lesbarkeit eines Textes [Gu52]                                                                         | <ul style="list-style-type: none"> <li>Anzahl der Wörter</li> <li>Anzahl der Sätze</li> <li>Anzahl der Wörter mit mehr als zwei Silben</li> </ul>                                                                        |
| Halstead-Metriken                | Verfahren zur Komplexitätsmessung von Softwarequelltexten, Beispiel: Kennzahl $D$ : Schwierigkeit, ein Programm zu verstehen [Ha77] | <ul style="list-style-type: none"> <li>Jeweils Anzahl unterschiedlicher Operanden und Operatoren</li> <li>Jeweils Gesamtanzahl von Operanden und Operatoren</li> </ul>                                                   |
| Wiener Sachtextformel (WSTF)     | Maßzahl zur Berechnung der Lesbarkeit von deutschsprachigen Texten [BV84]                                                           | <ul style="list-style-type: none"> <li>Anteil der Wörter mit mehr als zwei Silben</li> <li>mittlere Satzlänge</li> <li>Anteil der Wörter mit mehr als sechs Buchstaben</li> <li>Anteil der einsilbigen Wörter</li> </ul> |
| Zyklomatische Komplexität (ZK)   | Messung der Komplexität von Software, um Verständnis-kritische Stellen zu identifizieren, auch McCabe-Metrik genannt [Mc76]         | <ul style="list-style-type: none"> <li>Anzahl bedingter Anweisungen</li> <li>Anzahl der Unterfunktionen</li> </ul>                                                                                                       |

Tabelle 2 Untersuchte Metriken und ihre Eingangsgrößen

## 2.2 Das Komplexitätsmaß $m_{KI}$

Die oben beschriebenen Metriken wurden für längere Texte entwickelt. Eine Auswahlfrage besteht aber in der Regel nur aus etwa ein bis zwei Sätzen. Um das

Verhalten der Metriken auch bei Anwendung auf derartige Beispiele nachvollziehen zu können, wurden zunächst die Berechnungsvorschriften analysiert und eine Liste der möglichen Eingangsgrößen erstellt. Zusätzlich wurden für ca. 20 ausgewählte Beispielfragen mit jeweils einem besonders markant ausgeprägten Merkmal (z. B. viele Fremdwörter, kurze Fragen, lange Antwortoptionen) die zugehörigen Metrikwerte bestimmt. Die verschiedenen Berechnungsvorschriften ergeben auf eine Frage angewendet unterschiedliche Werte für die Verständlichkeit bzw. Lesbarkeit. So hat FRE einen typischen Wertebereich von 0-100, wobei 0 als sehr schwer gilt. Der Wert des GFI hingegen entspricht der jeweiligen Schulklassenstufe, in der diese Fertigkeit erreicht sein sollte. Um die verschiedenen Ergebnisse vergleichbar zu machen, wurden mit Hilfe einer Tabelle die verschiedenen Wertebereiche auf Klassen von 1 bis 6 abgebildet („Normierung“, s. Tabelle 3) Daraus konnte im Wesentlichen die Bedeutung der jeweiligen Eingangsgrößen abgeleitet werden. Mit diesem Wissen und dem Ziel einer möglichst einfachen und automatisierbaren Metrik wurde dann mit verschiedenen Varianten experimentiert.

|            |   | Metrik    |           |             |             |             |
|------------|---|-----------|-----------|-------------|-------------|-------------|
|            |   | Halstead  | ZK        | WSTF        | FRE         | GFI         |
| Normierung | 1 | bis 1,4   | bis 3,5   | bis 5,1     | über 65,6   | bis 2,11    |
|            | 2 | 1,5 - 2,4 | 3,6 – 5,0 | 5,2 – 8,4   | 65,6 – 43,9 | 2,12 – 3,11 |
|            | 3 | 2,5 - 3,4 | 5,1 – 6,5 | 8,5 – 11,7  | 43,8 – 22,2 | 3,12 – 4,11 |
|            | 4 | 3,5 - 4,4 | 6,6 – 8,0 | 11,8 – 15,0 | 22,1 – 0,5  | 4,12 – 5,11 |
|            | 5 | 4,5 - 5,4 | 8,1 – 9,5 | 15,1 – 18,3 | 0,4 – -21,3 | 5,12 – 6,11 |
|            | 6 | ab 5,5    | ab 9,6    | ab 18,4     | unter -21,3 | ab 6,12     |

Tabelle 3 Normierungstabelle der Metriken

Abbildung 1 zeigt eine Beispielfrage, die sehr hohe Werte bezüglich des Merkmals Fach- und Fremdwörter ( $x=3$  bei einem Mittelwert von 1,4) und langer Worte ( $x=13$ , Mittelwert: 7,0) hat. Die normierte Bewertung einer solchen Frage ist in Tabelle 4 zu sehen.

|                                                                                                                                   |                               |
|-----------------------------------------------------------------------------------------------------------------------------------|-------------------------------|
| Zu welchem umweltschädlichen Phänomen kann die Emission von Schwefeldioxid (z.B. bei industriellen Verbrennungsprozessen) führen? |                               |
| Saurer Regen                                                                                                                      | Verstärkung Treibhauseffektes |
| Extreme Wolkenbildung                                                                                                             | keine besondere Auswirkung    |

Abbildung 1: Beispielfrage mit vielen langen Worten und Fremdworten

| Metrik          | Halstead | ZK | WSTF  | FRE  | GFI  |
|-----------------|----------|----|-------|------|------|
| Wert            | 16       | 7  | 13,85 | 7,53 | 6,64 |
| Normierter Wert | 6        | 4  | 4     | 4    | 6    |

Tabelle 4: Werte und normierte Werte der Metriken für die Frage aus Abbildung 1

Die entworfene Metrik  $m_{Kt}$  enthält sieben Eingangsgrößen. Zu diesen weitestgehend objektiv feststellbaren Indikatoren zählen die Wortanzahl im Fragenstamm ( $W_F$ ) und in den Antwortoptionen ( $W_A$ ), die Anzahl der langen Wörter ( $W_L$ ), der Fremd- oder Fachwörter ( $W_{Fr}$ ) und der Satzzeichen ( $SZ$ ) sowie eine gleichmäßige Länge der Antwortoptionen ( $GL_A$ ) und die Nicht-Nutzung von Signalwörtern in den Antwortoptionen ( $SW$ ). Signalwörter sind Wörter wie „immer“ oder „nie“, die einer Antwort einen absoluten Anspruch geben, der jedoch gewöhnlich nicht erfüllbar ist und damit diese Antwort als richtige ausschließen [Po93].

Indikatoren, die nur mit Hilfe von Fachwissen bestimmbar sind, d.h. nicht oder nur sehr schwer automatisiert bestimmbar sind, wurden nicht weiter betrachtet. Darunter fallen u.a. die Bewertungen, ob es sich um Fang- oder Trickfragen handelt, wie sinnhaft die Distraktoren sind und ob die Frage auch ohne Vorgabe von Antwortmöglichkeiten zu beantworten ist. Die ausgewählten Indikatoren wurden in einer Berechnungsvorschrift zu einer Kennzahl, dem Maß für die textuelle Komplexität  $m_{Kt}$  miteinander verknüpft. Dabei wurde jedem Indikator gemäß einer Transformationstabelle ein ganzzahliger Wert zwischen 0 bis 3 zugewiesen. Tabelle 5 zeigt als Beispiel die Transformationstabelle für  $W_F$ . Eine Ausnahme ist der Indikator  $SW$ . Hier erfolgte eine dichotome Bewertung: 0 für die Abwesenheit von Signalwörtern, 3 falls welche vorhanden waren. Als langes Wort gilt eines ab drei Silben. Die Festlegung als Fremd- oder Fachwort kann über das Vorhandensein in einem (Fach-)Wörterbuch erfolgen, unterliegt jedoch auch subjektiven Einflüssen.

| Anzahl Wörter | Wert $W_F$ |
|---------------|------------|
| $\geq 13$     | 3          |
| 10-12         | 2          |
| 7-9           | 1          |
| $\leq 6$      | 0          |

Tabelle 5 Transformationstabelle Indikator *Wortanzahl im Fragenstamm* ( $W_F$ )

Formel 1 ist die hier genutzte Berechnungsvorschrift zur Bestimmung eines Maßes für die textuelle Komplexität  $m_{Kt}$ . Durch die den Wertebereich der sieben Summanden von 0 bis 3 kann der Wert dieser Metrik zwischen 0 und 21 schwanken. Dabei zeigen geringere Werte auch eine geringere Komplexität an.

$$m_{Kt} = W_F + W_A + W_L + W_{Fr} + SZ + GL_A + SW \quad (1)$$

### 2.3 Validierung

Diese Berechnungsvorschrift ist nicht als allgemein gültig zu sehen, sondern wurde im Rahmen dieser Betrachtungen entwickelt und hat sich für Beispielfragen als funktionierend erwiesen. Es sind jedoch auch weitere sinnvolle Alternativen denkbar, u.a. mit stetigen Transformationsfunktionen sowie zusätzliche Gewichtungen der einzelnen Summanden.

Zur Abklärung der Funktionalität der Berechnungsvorschrift wurde sie auf weitere Fragen angewendet. Zum einen waren dies Fragen zum Thema *Allgemeinwissen*. (vgl. Tabelle 1). Diese wurden aus dem beliebten entsprechenden Topic *Allgemeinwissen* der Quiz-App *QuizUp* extrahiert und sind daher als typische Unterhaltungsfragen zu bewerten. Ein weiterer Fragencorpus hatte das fachliche Thema *Anaerobtechnik*. Tabelle 6 zeigt die Ergebnisse der Metrik für diese Corpora. Es ist ersichtlich, dass die Metrik die vermutete unterschiedliche Komplexität von Fach- und Unterhaltungsfragen widerspiegelt. Mit 7,2 bzw. 7,8 ist der Wert für Fachfragen höher als für die Unterhaltungsfragen (4,8). Wesentliche Gründe hierfür sind die höhere Anzahl an Fremd- und Fachwörtern, langen Wörtern und längere Antwortoptionen. Bezüglich der gleichen Länge der Antwortoptionen besteht noch Optimierungspotenzial bei den

Fachfragen. Signalwörter sind hingegen kaum vorhanden. Hier scheinen die Richtlinien zur Erstellung von Auswahlfragen bereits in der Praxis angekommen zu sein.

| Corpus<br>Attribute/<br>Indikatoren | Allgemein-<br>wissen                     | Technische<br>Luftrein-<br>haltung | Anaerob-<br>technik<br>(orig.) | Anaerob-<br>technik<br>(rev.) |
|-------------------------------------|------------------------------------------|------------------------------------|--------------------------------|-------------------------------|
| QuizUp-Topic                        | <i>Lukas ‘<br/>Allgemeinwis-<br/>sen</i> | <i>Luftikusse</i>                  | <i>Anaerobier<br/>Stout</i>    | <i>Anaerobier<br/>Lager</i>   |
| Anzahl der Fragen                   | 109                                      | 162                                | 40                             | 40                            |
| $W_F$                               | 1,4                                      | 1,5                                | 1,3                            | 1,0                           |
| $W_A$                               | 0,7                                      | 1,1                                | 1,3                            | 0,7                           |
| $W_L$                               | 1,1                                      | 2,3                                | 2,6                            | 2,1                           |
| $W_{Fr}$                            | 0,4                                      | 1,1                                | 1,6                            | 1,1                           |
| $SZ$                                | 1,1                                      | 0,9                                | 0,5                            | 0,3                           |
| $GL_A$                              | 0,2                                      | 0,3                                | 0,4                            | 0,2                           |
| $SW$                                | 0,0                                      | 0,04                               | 0,2                            | 0,0                           |
| $m_{Kt}$                            | <b>4,8</b>                               | <b>7,2</b>                         | <b>7,8</b>                     | <b>5,4</b>                    |

Tabelle 6 Ergebnisse der Metrik  $m_{Kt}$  für verschiedene Fragencorpora

### 3 Die Anwendung der Metrik

#### 3.1 Überarbeitung des Fragenkatalogs

Den aktuellen Stand der Empfehlungen für Entwurfsrichtlinien haben wir in [SK16] zusammengefasst. Durch den Vergleich der strukturellen Charakteristiken von Fachfragen mit denen von Unterhaltungsfragen zeigt sich, dass diese Richtlinien im Wesentlichen berücksichtigt werden. Hauptsächlichste Ansatzpunkte für eine Optimierung sind eine Reduktion der langen Wörter sowie der Wortanzahl in den Antworten [HR13]. Dabei ist jedoch der Lernkontext zu beachten: essentielle Informationen sollten nicht eliminiert werden.

Um zu prüfen, ob eine Modifikation der Fragen unter dieser Maßgabe zu einer geringeren textuellen Komplexität und damit verbunden, zu einer besseren Nutzungsmöglichkeit im Rahmen einer Quiz-App führen kann, wurde der Fragencorpus zum Thema *Anaerobtechnik* diesbezüglich bearbeitet. Wie in Tabelle 6 zu sehen ist, konnte durch die Änderungen der Wert für die Metrik mit 5,4 deutlich – fast auf das Niveau der Unterhaltungsfragen - verbessert werden.

### 3.2 Nutzerstudie

Mit Hilfe einer Nutzerstudie wurde die Wirkung der Änderungen überprüft: 14 Studierende des Studiengangs Umweltingenieurwesens spielten in der Lehrveranstaltung *Anaerobtechnik* hintereinander die drei QuizUp-Themen *Anaerobier Stout* (originaler Fragenkatalog), *Anaerobier Lager* (überarbeiteter Fragenkatalog) und *Lukas' Allgemeinwissen* für jeweils 20 Minuten gegeneinander. Es wurden zwei Gruppen gebildet, die die beiden Anaerobtechnik-Fragenkataloge in jeweils umgekehrter Reihenfolge spielten, um die Auswirkungen möglicherweise durch die Reihenfolge auftretende Effekte auszuschließen. Den Probanden wurden die Unterschiede zwischen den beiden Anaerobtechnik-Fragenkatalogen nicht erläutert. Nach jeder Spielphase wurden Fragebögen ausgegeben, in denen die Probanden zu ihren Eindrücken befragt wurden. Es sollten dabei die folgenden Hypothesen überprüft werden:

H1: Die Überarbeitung des Fragenkatalogs hat seine Verständlichkeit erhöht.

H2: Der überarbeitete Fragenkatalog senkt den Zeitdruck.

H3: Die Überarbeitung des Fragenkatalogs führt zu einem höheren Spielspaß.

### 3.3 Ergebnisse

**H1.** Die **Verständlichkeit** des Fragenkatalogs wurde mit Hilfe der Fragen, ob die Auswahlfrage präzise gestellt und inhaltlich korrekt sei, bewertet. Dazu wurden jeweils 5-Punkt Likert-Skalen genutzt, wobei fünf Punkte hohe Zustimmung signalisierten. Zwischen den beiden Anaerobtechnik-Fragekatalogen gab es nur geringe Unterschiede (3,32 vs. 3,39). Hingegen war der Wert für die Unterhaltungsfragen mit 4,19 wesentlich höher. Mit Hilfe der Überarbeitung konnte die Verständlichkeit daher nicht erhöht werden.

**H2.** Der auftretende **Zeitdruck** wurde mit den drei Items „Es war schwierig, die Fragen in der zur Verfügung stehenden Zeit zu beantworten“, „Ich hatte Probleme, die Fragen in der zur Verfügung stehenden Zeit inhaltlich zu verstehen.“ und „Der Zeitdruck hat dafür gesorgt, dass ich nicht sorgfältig gelesen habe.“. Auch hier wiesen die Unterhaltungsfragen mit einem Durchschnittswert von 2,56 den besten Wert auf. Die überarbeiteten Fragen waren zwar mit 3,19 deutlich schlechter, verursachten jedoch auch merklich weniger Zeitdruck als die Originalfragen (3,64). Die Überarbeitung sorgte daher für weniger Zeitdruck.

**H3.** Der erreichte **Spielspaß** wurde mit Hilfe der In-Game-Variante des *Game Experience Questionnaire* (GEQ) [IKP13] überprüft. Die Ergebnisse sind nicht eindeutig (s. Abbildung 2). Zwar zeigen sich für positiv mit Spielspaß verbundene Merkmale wie *Competence*, *Challenge* und *Positive Affect* gleichmäßig gute Werte, jedoch sind diese Werte nicht einheitlich besser für einen der Fragenkataloge. Die Werte für *Negative Affect* und *Tension* hingegen sind für die beiden Fachfragenkataloge erkennbar

schlechter als für das Allgemeinwissen. Demgegenüber fühlen sich die Studierenden in den Fachfragen kompetenter als beim Allgemeinwissen. Insgesamt ergeben sich für alle drei Themen spieltypische Ausprägungen der Merkmale, jedoch liefern diese Werte keine Grundlage dafür, Hypothese 3 anzunehmen.

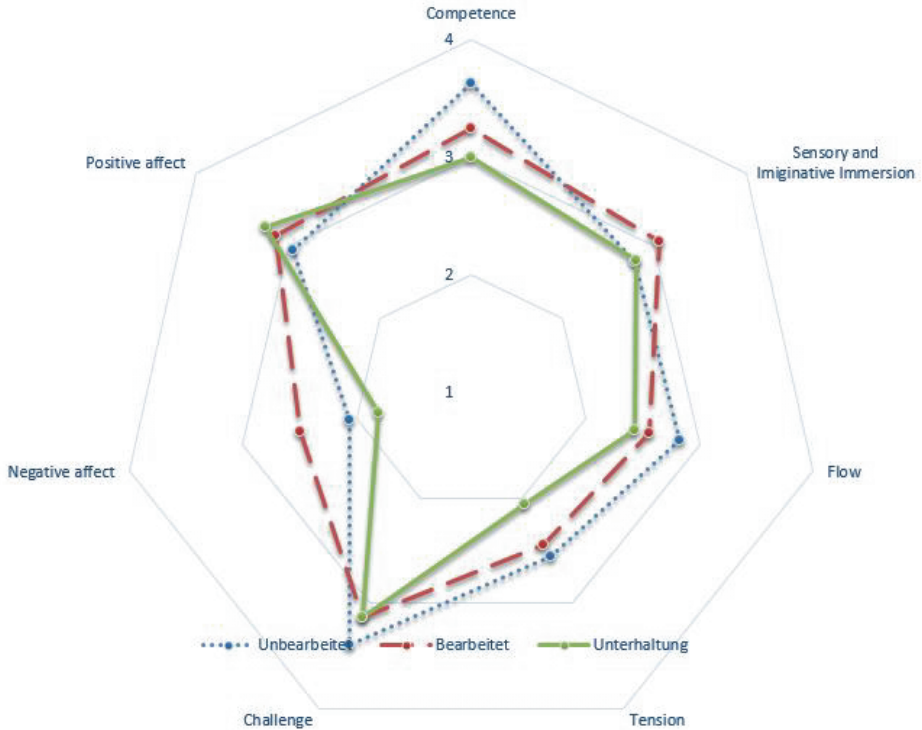


Abbildung 2 Ergebnisse GEQ

## 4 Diskussion und Schlussfolgerungen

Ziel der Arbeit war es, den offensichtlichen Popularitäts-Unterschied zwischen Fach- und Unterhaltungsfragen im Rahmen von Quiz-Apps aufzuarbeiten und insbesondere durch eine Analyse der textuellen Struktur zu einer höheren Attraktivität von Fachfragen beizutragen. Dazu wurde eine Metrik zur Messung der Komplexität einer Auswahlfrage entworfen, die weitestgehend automatisiert auf die Fragen angewendet werden könnte. Beim Entwurf der Metrik wurden u.a. Lesbarkeitsformeln genutzt um Hinweise auf die Optimierung der Fragen zu geben. Mit der Metrik konnten zum einen die strukturellen Unterschiede zwischen den beiden Fragenkategorien dokumentiert werden. Zum anderen wurden daraus Entwurfsgrundsätze erweitert, die insbesondere das Augenmerk auf

Gesamtlänge sowie Anzahl der Fach- und Fremdwörter legen. Ein diesbezüglich modifizierter Fragenkatalog erzielte beim Spiel in einer Quiz-App bessere Werte für den empfundenen Zeitdruck als der ursprüngliche Katalog. Die Werte eines Unterhaltungsthemas konnten jedoch nicht erreicht werden.

Dieses Ergebnis unterliegt sicherlich einer Reihe von Einschränkungen und Unsicherheiten. Die Stichprobe ist sehr klein, sodass bisher ausschließlich deskriptive Auswertungen möglich waren. Außerdem reichten die durchgeführten Fragenänderungen nicht aus, um bessere Ergebnisse in Bezug auf Spielspaß und Verständlichkeit zu erzielen. Während der Spielspaß mit der experimentellen Atmosphäre begründet werden könnte, deutet die Verständlichkeit eher auf grundsätzliche Spannungsfelder beim Fragenentwurf hin: durch eine versuchte Reduktion der Länge sowie der Fach- und Fremdwörter könnte die notwendige fachliche Präzision möglicherweise eingeschränkt werden. Inwieweit es möglich ist, in kurzen und prägnanten Worten einen für das Lernziel relevanten Sachverhalt auszudrücken, muss noch weiter evaluiert werden. So ist es zwar gelungen, die Fachfragen fast auf das Verständlichkeits-Niveau der Unterhaltungsfragen anzugleichen. Jedoch ist aufgrund des komplexeren Fachwissens nicht mit derselben Lesbarkeit zu rechnen. Gleichfalls ist zu prüfen, inwieweit durch die Vereinfachung auch ein Teil des Wissens verloren geht. Möglicherweise ist es notwendig, die Fragen in mehrere Fragen zu unterteilen, um dieselben Lernziele erreichen zu können, dabei aber durch einen schnelleren Lesefluss in der Quiz-Applikation einen höheren Spielempfinden zu erzeugen. Es ist einschränkend anzumerken, dass textuelle Komplexität nur eine von mehreren Charakteristiken ist, die für die Attraktivität der Fragen im Quiz-App-Kontext von Bedeutung sind. Zu den weiteren Faktoren zählen u.a. die Spielbereitschaft der Lernenden sowie die Inhalte der Frage selbst. Die Metrik als solche hat sich zwar als funktionierend erwiesen. Jedoch kann zwischen den Indikatoren der Anzahl von Fach- und Fremdwörtern sowie der Anzahl der langen Wörter ein Zusammenhang vermutet werden. Dies ist als Indiz zu werten, dass eine Überarbeitung der Metrik ihre Vereinfachung nach sich ziehen könnte. In einer folgenden Studie sollten der Einfluss textueller Komplexität bei Fach- und Unterhaltungsthemen auf den Lernerfolg sowie die Akzeptanz beider Themenbereiche untersucht werden.

Die dargestellte Metrik in Verbindung mit der Validierung hat jedoch einen Arbeitsstand erreicht, mit dem eine analytische Qualitätssicherung bei der Fragerstellung stattfinden kann. Durch eine Implementierung in einem Frageneditor könnten Hinweise zur Steigerung der Lesbarkeit direkt bei der Fragerstellung gegeben werden.

## Literaturverzeichnis

- [Am78] Amstad, T.: Wie verständlich sind unsere Zeitungen? Studenten-Schreib-Service, 1978.
- [BR07] Butler, A. C.; Roediger, H. L.: Testing improves long-term retention in a

- simulated classroom setting. *European Journal of Cognitive Psychology* 19, S. 514–527, 2007.
- [BV84] Bamberger, R.; Vanecek, E.: Lesen, verstehen, lernen, schreiben: die Schwierigkeitsstufen von Texten in deutscher Sprache. Jugend und Volk, 1984.
- [Fl48] Flesch, R.: A new readability yardstick. *Journal of Applied Psychology* 32, S. 221, 1948.
- [Gu52] Gunning, R.: *The Technique of Clear Writing*. McGraw-Hill, New York, 1952.
- [Ha14] Hardinghaus, B.: Interview mit LeBernd. Bernd Schneider über Quizduell, <http://www.spiegel.de/netzwelt/games/interview-mit-lebernd-bernd-schneider-ueber-quizduell-a-951529.html>, Stand: 11.02.2015.
- [Ha77] Halstead, M. H.: *Elements of Software Science (Operating and Programming Systems Series)*. Elsevier Science Inc, New York, NY, USA, 1977.
- [HR13] Haladyna, T. M.; Rodriguez, M. C.: *Developing and Validating Test Items*. Routledge, New York, 2013.
- [IKP13] IJsselsteijn, W. A.; Kort, Y. A. W. de; Poels, K.: The Game Experience Questionnaire: Development of a self-report measure to assess the psychological impact of digital games. Manuscript in Preparation, 2013.
- [Mc76] McCabe, T. J.: *A Complexity Measure*, 1976.
- [Pl14] Plain Vanilla: QuizUp - Connecting people through shared interests, <https://www.quizup.com/>, Stand: 12.01.2016.
- [Po93] Poundstone, W.: *Prisoner's Dilemma/John von Neumann, Game Theory and the Puzzle of the Bomb*. Anchor, 1993.
- [Ru14] Russolillo, S.: QuizUp: The Next 'It' Game App?, <http://www.palmbeachpost.com/videos/news/is-quizup-the-next-it-game-app/vCYDgf/> <http://live.wsj.com/video/quizup-the-next-it-game-app/>, Stand: 28.01.2016.
- [SK16] Söbke, H.; Kämmerer, F.: Vermessene Fragen: Metriken als Ansatz automatisierter analytischer und konstruktiver Qualitätssicherung von Mehrfachauswahlfragen für mobile digitale Medien. In (Pfau, W. et al. Hrsg.): *Teaching Trends 2016-Vielfalt in der Lehre*, Clausthal-Zellerfeld, 10./11. November 2016. Waxmann, Münster, S. 153–162, 2016.
- [Sö15] Söbke, H.: Space for seriousness? Player Behavior and Motivation in Quiz Apps. In (Chorianopoulos, K.; al, E. Hrsg.): *Entertainment Computing – ICEC 2015 14th International Conference*, ICEC 2015 Trondheim,

Norway, September 29 – October 2, 2015 Proceedings. Springer, Cham, S. 482-489, 2015.

- [SW16] Söbke, H.; Weitze, L.: 2. The Challenge to Nurture Challenge Students' Perception of a Commercial Quiz App as a Learning Tool. In (Wallner, G. et al. Hrsg.): Entertainment Computing - ICEC 2016 - 15th International Conference, Vienna, Austria, September 28-30, 2016, Proceedings. Springer International Publishing, S. 15-23, 2016.