Methoden zur Messung und Modellierung des Zufalls in menschengenerierten Zahlensequenzen

Inaugural-Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Tim Angelike

aus Erkelenz

Düsseldorf, Dezember 2023

aus dem Institut für Experimentelle Psychologie

der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit Genehmigung der

Mathematisch-Naturwissenschaftlichen Fakultät der

Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Jochen Musch

2. Prof. Dr. Gerhard Jocham

Tag der mündlichen Prüfung: 20.02.2024

Danksagung

Ich möchte mich an dieser Stelle bei Prof. Dr. Jochen Musch herzlich für die Betreuung dieser Promotion sowie für das konstruktive und kritische Feedback bedanken. Außerdem möchte ich mich bei Prof. Dr. Gerhard Jocham für die Zweitbegutachtung dieser Arbeit bedanken.

Mein Dank gilt all denjenigen, die mich über die letzten beiden Jahre hinweg begleitet haben. Hier zu nennen sind insbesondere meine Eltern und meine Schwester. Vielen Dank für Eure Unterstützung und Euer Interesse an meinem Forschungsthema! Besonderer Dank gilt auch der „Mensagruppe“, die mich fast jeden Tag der letzten beiden Jahre begleitet hat. Mir ist bewusst, dass ich Euch viel von Zufallszahlengenerierung erzählt habe und Ihr mir dennoch (meistens auch geduldig) zugehört habt. Ein ähnliches Schicksal widerfuhr insbesondere Martin und Katharina, denen ich ebenfalls für ihr offenes Ohr und ihre kritischen Rückmeldungen danke.

Inhaltsverzeichnis

Zusammenfassung

In verschiedenen Feldern psychologischer Forschung wird untersucht, warum Menschen Schwierigkeiten damit haben, sich zufällig zu verhalten. Dies wird oft mit der sogenannten Zufallszahlengenerierungsaufgabe (Random Number Generation task; RNG task) untersucht. Bei dieser Aufgabe werden Teilnehmende dazu aufgefordert, eine möglichst zufällige Abfolge von Zahlen zu generieren. Dabei machen Menschen typischerweise Fehler, die zu wenig zufälligen und stark vorhersehbaren Sequenzen führen. So meiden Menschen beispielsweise direkte Wiederholungen einer Zahl, wiederholen bestimmte Zahlenabfolgen häufiger als andere und achten zu sehr darauf, möglichst schnell alle zur Auswahl stehenden Zahlen zu verwenden; sie wählen deshalb Zahlen mit erhöhter Wahrscheinlichkeit, die zuletzt lange nicht gewählt wurden. Traditionellerweise werden in der psychologischen Forschung Maße zur Quantifizierung von Zufälligkeit eingesetzt, die unmittelbar typische Fehler bei der Generierung von Zufall widerspiegeln. In anderen Forschungsdisziplinen wie der Informatik werden eher Maße für Zufälligkeit verwendet, die auf einer stärkeren mathematischen Grundlage beruhen. Hierzu gehören insbesondere Methoden zur Berechnung der Entropie und der algorithmischen Komplexität. Neu entwickelte Maße zur Approximation algorithmischer Komplexität versprechen Regelmäßigkeiten jeder Art in einer Sequenz zu detektieren und wurden kürzlich für eine Anwendung in der psychologischen Forschung vorgeschlagen. Es fehlt insgesamt jedoch ein klarer Konsens darüber, welche Methoden zur Messung von Zufälligkeit am besten geeignet sind, um menschengenerierte Sequenzen von Zahlen zu analysieren. In der ersten Untersuchung der vorliegenden Arbeit wird die erste umfassende, vergleichende Untersuchung unterschiedlicher Methoden zur Quantifizierung von Zufälligkeit vorgenommen. Dabei wird ein neuer klassifikationsbasierter Validierungsansatz verfolgt, der prüft, wie gut verschiedene Maße für

Zufälligkeit zwischen menschengenerierten und echt zufälligen Sequenzen zu differenzieren vermögen. Dazu wird untersucht, wie sensitiv die untersuchten Maße für typisch menschliche Abweichungen von echt zufälligem Verhalten sind. Die Ergebnisse dieser Untersuchung zeigen, dass insbesondere Maße für algorithmische Komplexität, aber auch einige der traditionell in der psychologischen Forschung angewandten Maße zur Quantifizierung der Abstände zwischen Wiederholungen von Zahlen gut zwischen menschengenerierten und zufälligen Sequenzen differenzieren können. Darüber hinaus werden Maße für Zufälligkeit identifiziert, die weniger gut dazu geeignet sind, menschengenerierte von zufälligen Sequenzen zu unterscheiden. In einer zweiten Untersuchung wird ein Ansatz zur formalen Modellierung des Verhaltens bei der Generierung zufälliger Sequenzen modifiziert und erweitert. Das hierfür neu vorgeschlagene Modell beinhaltet einen Wiederholungs-, einen Distanz- und einen Zirkulationsparameter. Der Wiederholungsparameter misst die Tendenz zu übermäßig vielen oder wenigen Wiederholungen. Der Distanzparameter steht für die Tendenz, bei aufeinander folgenden Wahlen entweder nah beieinander oder weit voneinander entfernt liegende Zahlenpaare zu generieren. Der Zirkulationsparameter reflektiert die Neigung, bevorzugt solche Zahlen zu generieren, die länger nicht verwendet wurden. Das Modell erlaubt eine gleichzeitige Schätzung dieser systematischen Fehler bei der Generierung zufälliger Sequenzen. Umfangreiche Computersimulationen auf der Basis der geschätzten Modellparameter bestätigen, dass mit dem hier erweiterten Modell menschliches Verhalten bei der Generierung von Zufallszahlen besser beschrieben und typische menschliche Besonderheiten und Fehler beim Generieren möglichst zufälliger Sequenzen auch besser als mit den bislang vorgeschlagenen Modellen vorhergesagt werden können.

Abstract

Several fields of psychological research have investigated the reasons why people have difficulty behaving randomly. This is often investigated with the so-called random number generation (RNG) task. In this task, participants are asked to generate a maximally random sequence of numbers. However, people typically make several errors in this task, resulting in sequences that are not random and highly predictable. For example, people avoid direct repetitions of a number, repeat certain subsequences of numbers more often than others, and cycle through all possible numbers in a sequence too quickly. The latter describes tendencies that lead to an increased probability of generating numbers that have not been generated in the recent choice history. Traditionally, psychological research investigating the ability to behave randomly has used measures to quantify randomness that directly reflect the assumed biases underlying the generation of random sequences. In contrast, other research disciplines, such as computer science, have proposed measures of randomness with a stronger mathematical foundation. Methods for computing entropy and algorithmic complexity are worth mentioning. In particular, recently developed measures for approximating algorithmic complexity, which promise to detect regularities and patterns of all kinds, have been proposed for application in psychological research. However, there is a general lack of consensus about which methods for measuring randomness are most appropriate for analyzing human-generated sequences of numbers. The first study in this dissertation was the first comprehensive, comparative investigation of different methods for quantifying randomness. It used a novel classification-based validation approach that focused on how well the measures of randomness examined were able to discriminate between human-generated and truly random sequences. This allowed the sensitivity of the randomness measures to typical human behavior in RNG tasks to be investigated. This study showed that measures of

algorithmic complexity, but also some measures traditionally used in psychological research to quantify the gaps between number repetitions, are particularly good at discriminating between human-generated and random sequences. In addition, this dissertation identified measures of randomness that appear to be less suitable for analyzing human-generated sequences. In a second investigation in this dissertation, a formal model of the systematic behaviors underlying the generation of random sequences in humans was modified and extended. The new model proposed here includes a repetition parameter, a distance parameter, and a cycling parameter. The repetition parameter expresses systematic tendencies toward too many or too few repetitions. The distance parameter allows the assessment of systematic behavior manifested in the generation of subsequent numbers that are either adjacent or distant to each other. The cycling parameter describes the tendency to generate numbers that have not been generated recently. This approach allows the simultaneous estimation of different systematic biases that people show when trying to generate random sequences. This study showed that the newly proposed model is better able to describe human behavior than the original model. Extensive computer simulations based on the estimated model parameters confirmed that the extended model can be used to better explain human biases when trying to generate random sequences.

1 Einleitung

1.1 Zur Rolle des Zufalls in der Psychologie

Menschen haben Schwierigkeiten, sich zufällig zu verhalten (Figurska et al., 2008; Ginsburg & Karpiuk, 1994). Beim Versuch dies zu tun, zeigen sie eine Vielzahl systematischer Verhaltensweisen, die es leicht machen, die Besonderheiten ihres Verhaltens zu charakterisieren (Towse & Neil, 1998) und vorherzusagen (Schulz et al., 2021; Shteingart & Loewenstein, 2016). Die Fähigkeit, sich zufällig und unvorhersehbar zu verhalten, wird häufig in sogenannten *Random Number Generation* (RNG)-Aufgaben untersucht. Obwohl diese Fähigkeit bereits seit den 1960er-Jahren erforscht wird (Baddeley, 1966), gibt es keinen klaren Konsens bezüglich der Frage, wie Zufälligkeit in menschengenerierten Sequenzen am besten quantifiziert werden sollte (Barbasz et al., 2008; Gauvrit et al., 2016; Wagenaar, 1972). Ziel dieser Dissertationsschrift ist es, einen umfassenden empirischen Vergleich zwischen bestehenden Methoden vorzunehmen, mit Hilfe derer man Zufälligkeit in menschengenerierten Sequenzen quantifizieren kann, um anschließend daraus Empfehlungen abzuleiten, welche Maße für Zufälligkeit besonders geeignet sind, um menschliches Verhalten zu analysieren und von echt zufälligem Verhalten zu unterscheiden. In einem zweiten Schritt wird ein kürzlich vorgeschlagener formaler Ansatz zur Modellierung menschlichen Verhaltens in RNG-Aufgaben erweitert und verbessert. Das verbesserte Modell erlaubt es, die zugrundeliegenden Systematiken bei der Generierung zufälliger Sequenzen genauer zu modellieren und mit höherer Treffergenauigkeit vorherzusagen, als dies bislang möglich war.

In RNG-Aufgaben werden Teilnehmende eines Experiments dazu aufgefordert, eine möglichst zufällige Reihe von Zahlen, die aus einem diskreten Intervall stammen, zu generieren (Ginsburg & Karpiuk, 1994; Towse & Neil, 1998). In der gängigsten Variante dieser Aufgabe dürfen bei der Generierung der Sequenzen die Zahlen von 1 bis 9 verwendet werden (Capone et

al., 2014; Jokar & Mikaili, 2012; Miyake et al., 2000, 2001; Schulz et al., 2012, 2021). Es gibt jedoch auch davon abweichende Formate, bei denen beispielsweise nur zwei mögliche Zahlen (Biesaga et al., 2021; Biesaga & Nowak, 2022; Gauvrit et al., 2016) oder sogar zehn oder mehr Zahlen verwendet werden dürfen (Towse, 1998; Towse & Cheshire, 2007). Näher betrachtet wird in dieser Dissertationsschrift das gängigste Aufgabenformat, in dem die Zahlen von 1 bis 9 benutzt werden sollen. Vor Durchführung einer RNG-Aufgabe werden Teilnehmende ausführlich über das Konzept und die Eigenschaften von Zufall aufgeklärt (Schulz et al., 2021; Towse & Cheshire, 2007). Ziel dieser Aufklärung ist die Messung der tatsächlichen Fähigkeit von Menschen, Zufall zu generieren, um nicht stattdessen fälschliche Annahmen über die Eigenschaften zufälliger Sequenzen zu untersuchen. Wenn Menschen versuchen, möglichst zufällige Sequenzen von Zahlen in einer RNG-Aufgabe zu generieren, zeigen sie dabei stark systematisches Verhalten. So meiden Menschen direkte Wiederholungen einer Zahl, weshalb es unmittelbar nach Generierung einer Zahl äußerst unwahrscheinlich ist, dass diese direkt wiederholt wird (Cooper, 2016; Towse, 1998). Menschen zirkulieren außerdem zu schnell durch die Menge aller verfügbaren Zahlen, was dazu führt, dass Zahlen, die länger nicht mehr verwendet wurden, im Vergleich zu Zahlen, die erst kürzlich generiert wurden, mit erhöhter Wahrscheinlichkeit für die Generierung der nächsten Zahl in einer Sequenz verwendet werden (Ginsburg & Karpiuk, 1994; Peters et al., 2007). Diese Verhaltensweise wurde in anderen Kontexten als *gambler's fallacy* (Tversky & Kahneman, 1971) untersucht. Bezeichnet wird damit die Neigung von Menschen, beim Glücksspiel Zahlen eine höhere Auftretenswahrscheinlichkeit beizumessen, die länger nicht mehr generiert wurden. Menschen tendieren außerdem auch zur Wiederholung interindividuell unterschiedlicher aber stereotyper Zahlenabfolgen, wie bestimmten Paaren oder Triplets, was ihr Verhalten ebenfalls vorhersehbar macht (Jokar & Mikaili, 2012; Schulz et al., 2012, 2021). Es lässt sich also

konstatieren, dass Menschen erhebliche Probleme damit haben, bewusst wirklich zufällige Zahlensequenzen zu generieren (Figurska et al., 2008; Ginsburg & Karpiuk, 1994; Towse & Neil, 1998).

Das experimentelle Paradigma der RNG-Aufgabe kann dazu verwendet werden, verschiedene kognitive Funktionen des Menschen zu erforschen. So zeigt die bisherige Forschung, dass beim Versuch, zufällige Sequenzen von Zahlen zu generieren, mehrere grundlegende kognitive Prozesse erforderlich sind. Es werden Gedächtnisprozesse benötigt, um sich während der RNG-Aufgabe daran zu erinnern, welche Zahlen zuletzt generiert wurden; es wird Aufmerksamkeit benötigt, um diese zuvor generierten Zahlen hinsichtlich möglicher Regelmäßigkeiten zu analysieren; und es wird Inhibitionsfähigkeit benötigt, um die Generierung solcher Zahlenabfolgen zu unterdrücken, die sonst zu systematischen Sequenzen von Zahlen führen würden (Cooper, 2016; Friedman & Miyake, 2004; Miyake et al., 2000). Da RNG-Aufgaben verschiedene kognitive Prozesse erfordern, werden sie auch gerne als Zweitaufgabe zur Ablenkung in anderen psychologischen Forschungsparadigmen eingesetzt (Howarth et al., 2016; Knott & Dewhurst, 2007; Miyake et al., 2001). Solche Studien zeigen, dass die gleichzeitige Bearbeitung einer RNG-Aufgabe zu einer Reduktion der Leistung in der Hauptaufgabe und in der RNG-Aufgabe führt. Dies unterstreicht, dass es sich beim Generieren von zufälligen Sequenzen um eine anstrengende Aufgabe handelt, die unterschiedliche kognitive Prozesse erfordert.

Dieser Befund wird dadurch bestärkt, dass die Zufälligkeit der generierten Sequenzen in einer RNG-Aufgabe von verschiedenen Faktoren abhängig ist, die in Zusammenhang mit kognitiven Funktionen gebracht werden, wie dem Alter der Teilnehmenden (Gauvrit et al., 2017; Heuer et al., 2010), dem Erholungszustand (Heuer et al., 2005) oder dem Niveau externer Ablenkung durch Töne und Geräusche (Marsh et al., 2013). RNG-Aufgaben finden auch in der

klinischen und neurowissenschaftlichen Psychologie eine Anwendung. Menschen mit verschiedenen neurologischen und psychiatrischen Erkrankungen schneiden in diesem experimentellen Paradigma schlechter ab als gesunde Personen (Gauvrit et al., 2016). Beispiele hierfür sind Personen mit Hirnverletzungen (Maes et al., 2011), Schizophrenie (Peters et al., 2007; Shinba et al., 2000), Parkinson (Williams et al., 2020) oder Demenz (Brugger et al., 1996).

1.2 Ansätze zur Messung von Zufälligkeit

Da RNG-Aufgaben in verschiedenen Forschungsbereichen der Psychologie verwendet werden, ist es wichtig ist, Zufälligkeit in menschengenerierten Sequenzen präzise messen zu können. Bei der Messung von Zufälligkeit besteht jedoch ein grundlegendes Problem: Betrachtet man eine gegebene Sequenz von Zahlen, ist es unmöglich mit Gewissheit zu sagen, ob diese durch Zufall oder durch einen anderen nicht-zufälligen Prozess generiert wurde. Ursächlich hierfür ist, dass unter Annahme eines zufälligen Prozesses als generierende Quelle jede Sequenz von Zahlen dieselbe Auftretenswahrscheinlichkeit hat (Gauvrit et al., 2016). Eine Sequenz mit einer stark repetitiven Struktur wie 1-2-1-2-1-2-1-2 hat unter Annahme eines zufälligen Prozesses dieselbe Auftretenswahrscheinlichkeit wie eine scheinbar weniger systematische Sequenz wie 1-1-2-2-2-1-1-2. Dies liegt darin begründet, dass jede Zahl in einer Sequenz immer dieselbe Auftretenswahrscheinlichkeit hat, als nächstes generiert zu werden, unabhängig davon, welche Zahlen zuvor generiert wurden. Ob eine Sequenz durch einen zufälligen Prozess generiert wurde, lässt sich folglich nie abschließend belegen oder widerlegen. Man würde jedoch aus der Tatsache, dass der ersten Beispielsequenz eine klare Systematik zugrunde liegt, schlussfolgern, dass diese mit einer geringeren Wahrscheinlichkeit von einem zufälligen Prozess generiert wurde als die zweite Beispielsequenz. Um die Zufälligkeit von Sequenzen zu quantifizieren, sucht man deshalb nach Hinweisen dafür, dass diese nicht vom Zufall, sondern durch einen Prozess generiert wurde, dem eine Systematik zugrunde liegt (Ginsburg & Karpiuk, 1994; Towse & Neil, 1998). Dabei werden verschiedene theoretische Ansätze zur Quantifizierung von Zufälligkeit in menschengenerierten Sequenzen verfolgt. Ein Vergleich dieser Ansätze wird in den folgenden Abschnitten vorgenommen.

In der psychologischen Forschung werden oft Maße für Zufälligkeit eingesetzt, die auf den Arbeiten von Ginsburg und Karpiuk (1994) sowie Towse und Neil (1998) aufbauen. In beiden Publikationen wurden eine Reihe unterschiedlicher Maße vorgeschlagen, um menschengenerierte Sequenzen hinsichtlich ihrer Zufälligkeit zu analysieren. Die Sammlung von 16 Maßen nach Towse und Neil stellt dabei eine Erweiterung und Verbesserung der Maße nach Ginsburg und Karpiuk dar. Für eine genaue Auflistung und Beschreibung dieser Maße wird auf die erste hier präsentierte Forschungsarbeit im Anhang verwiesen. Gemeinsam ist den vorgeschlagenen Maßen für Zufälligkeit, dass mit ihnen spezifische Systematiken quantifiziert werden, die menschengenerierte Sequenzen häufig aufweisen. Beispiele hierfür sind Maße, die angeben, wie häufig es in einer Sequenz zu Wiederholungen kommt; Maße wie der Coupon-Score, die angeben, wie viele Zahlen im Durchschnitt benötigt werden, bis alle möglichen Zahlen in einer Sequenz mindestens einmal verwendet wurden; und Maße, die quantifizieren, ob eine Person eine Tendenz dazu hat, bestimmte Wertpaare mit erhöhter Wahrscheinlichkeit zu wiederholen.

Aufgrund der hohen Zahl derartiger Maße werden diese oft mithilfe von Hauptkomponentenanalysen aggregiert, um die Interpretation dieser Maße zu erleichtern (Maes et al., 2011; Oomens et al., 2023; Peters et al., 2007). Grundlage solcher Analysen ist dann die Matrix der über alle Teilnehmenden berechneten Korrelationen zwischen den verwendeten Zufälligkeitsmaßen. Ziel der Hauptkomponentenanalyse ist es, die Daten (in diesem Fall die Zufälligkeitsmaße) auf die wesentlichen zugrundeliegenden Dimensionen zu reduzieren, dabei jedoch möglichst viel Information aus den ursprünglichen Daten zu erhalten, um sie leichter interpretieren zu können (Ginsburg & Karpiuk, 1994; Towse & Neil, 1998). Dies erweist sich insofern als notwendig, als die verschiedenen Zufälligkeitsmaße in den vorgeschlagenen Sammlungen oft stark miteinander korreliert sind, sodass diese nicht unabhängig voneinander

interpretiert werden können. Eine Hauptkomponentenanalyse verspricht eine Identifikation aller wesentlichen, den Zufälligkeitsmaßen zugrunde liegenden Komponenten (Oomens et al., 2023). Die Ergebnisse einer Hauptkomponentenanalyse über die Sammlung von Maßen nach Ginsburg und Karpiuk führten zur Extraktion der drei Faktoren Zirkulation, Serienbildung und Wiederholungen (Englisch: cycling, seriation, repetition; Ginsburg & Karpiuk, 1994). Als „Zirkulation“ beschreibt man Verhalten, bei dem Menschen versuchen, alle möglichen Zahlen gleich häufig zu verwenden, was dazu führt, dass zu schnell durch alle möglichen Zahlen durchrotiert wird. Die Komponente „Serienbildung“ beschreibt, dass Menschen bevorzugt bestimmte Abfolgen von Zahlen generieren (beispielsweise auf- oder absteigende Zahlensequenzen). Die dritte Komponente „Wiederholung“ spiegelt wider, ob eine Tendenz zu Wiederholungen oder zur Vermeidung von Wiederholungen besteht. Towse und Neil (1998) führten über die von ihnen vorgeschlagenen Maße ebenfalls eine Hauptkomponentenanalyse durch und erhielten damit vier Hauptkomponenten. Eine dieser Komponenten ähnelt inhaltlich der Zirkulationskomponente nach Ginsburg und Karpiuk, eine weitere ähnelt der Serienbildung, und die beiden verbleibenden Komponenten nach Towse und Neil unterteilen die Wiederholungskomponente nach Ginsburg und Karpiuk in zwei Untergruppen bezüglich Wiederholungen, die in einer Sequenz entweder mit einem geringen oder einem langen Abstand zueinander auftreten.

Einige der in den Sammlungen nach Ginsburg und Karpiuk (1994) und Towse und Neil (1998) enthaltenen Maße haben ihren Ursprung in der Informationstheorie und sind Varianten der klassischen Shannon-Entropie (Shannon, 1948). Mit der Entropie kann quantifiziert werden, ob eine Gleichverteilung in der Verwendung aller Zahlen in einer Sequenz besteht. Maximale Werte für Entropie werden erzielt, wenn in einer Sequenz alle möglichen Zahlen genau gleich häufig

auftreten. Die minimale Entropie (= 0) wird erzielt, wenn eine Sequenz nur aus der Wiederholung einer einzigen Zahl besteht. Das Konzept der Entropie lässt sich erweitern auf beliebig große Zahlenblöcke, was es erlaubt zu quantifizieren, ob beispielsweise Wertpaare aufeinanderfolgender Zahlen besonders häufig auftreten. Diese Erweiterung der klassischen Entropie bezeichnet man auch als Blockentropie (Moore et al., 2018; Shannon, 1948).

Ein alternativer Ansatz zur Quantifizierung von Zufälligkeit entstammt der algorithmischen Komplexitätstheorie (Gauvrit et al., 2014; Soler-Toscano et al., 2014). In diesem Ansatz wird die Komplexität einer Sequenz definiert als die Länge des kürzesten Computerprogramms, die eine gegebene Sequenz produzieren kann (Gauvrit et al., 2016). Sequenzen mit einer simplen Struktur (wie das zuvor gewählte Beispiel 1-2-1-2-1-2-1-2) können durch ein relativ kurzes Computerprogramm generiert werden (beispielsweise „wiederhole die Zahlensequenz 1-2 vier Mal“), wohingegen komplexe Sequenzen nicht auf einfache Programm-anweisungen reduziert werden können. Der Begriff der algorithmischen Komplexität ist eng verknüpft mit dem Konzept von Zufall, da eine Sequenz mit maximaler Komplexität keine Regelmäßigkeiten und systematischen Abfolgen besitzt, die man durch ein Computerprogramm abkürzen könnte (Zenil et al., 2018). Eine komplexe Sequenz ist demnach eine Sequenz, die sich durch eine Abwesenheit von Regelmäßigkeiten auszeichnet und bei der man auf Basis vorangegangener Elemente in einer Sequenz nicht vorhersagen kann, welche Zahl als nächstes auftritt. Dieser Ansatz hat den Vorteil, dass er nicht auf der Analyse spezifischer statistischer Regelmäßigkeiten basiert wie die zuvor beschrieben Ansätze (Gauvrit et al., 2016). So wäre bei ausschließlicher Betrachtung binärer Sequenzen die klassische Shannon-Entropie für die Beispielsequenz 1-2-1-2-1-2-1-2 maximal, da beide Werte 1 und 2 genau gleich häufig auftreten.

Ansätze algorithmischer Komplexität hingegen versprechen, solche Regelmäßigkeiten detektieren zu können.

Die algorithmische Komplexität kann allerdings nicht direkt berechnet werden und wird deshalb typischerweise approximiert durch Kompressionsalgorithmen für längere Sequenzen (Gauvrit et al., 2016; Soler-Toscano et al., 2014; Zenil et al., 2018), wobei der hier zugrundeliegende Gedanke ist, dass zufällige Sequenzen weniger gut komprimierbar sind als Sequenzen, die eine starke zugrunde liegende Systematik aufweisen. Dies ist jedoch für die vergleichsweise kurzen menschengenerierten Sequenzen von oft nur 100 Zahlen oder weniger problematisch, da der so approximierte Wert für die algorithmische Komplexität stark vom verwendeten Kompressionsalgorithmus abhängig ist (Gauvrit et al., 2016; Zenil et al., 2018). So gibt es verschiedene Varianten von Kompressionsalgorithmen, die sich zwar ähneln, jedoch nicht identisch sind in ihrer Funktionsweise (Lempel & Ziv, 1976; Ziv & Lempel, 1977). Diese Implementierungsunterschiede können sich insbesondere bei der Anwendung auf kurze Sequenzen bemerkbar machen.

Kürzlich wurde ein neuer Ansatz zur Approximation der algorithmischen Komplexität vorgeschlagen. Der neu vorgeschlagene Ansatz ist explizit auf die Analyse kurzer Sequenzen von maximal 10 bis 12 Zeichen ausgelegt (Gauvrit et al., 2014, 2016; Soler-Toscano et al., 2014). Erfolgsversprechend ist dieser Ansatz insofern, als lediglich die Abwesenheit von Regelmäßigkeiten in einer Sequenz als Definition von Zufall zugrunde gelegt wird. Diese Definition ist unabhängig von spezifischen statistischen Regelmäßigkeiten wie beispielsweise dem gehäuften Auftreten von Wiederholungen oder Wertpaaren in einer Sequenz, die den typischerweise in der psychologischen Forschung verfolgten Ansatz kennzeichnen. Außerdem ist bisher kein anderes Maß für Zufälligkeit explizit auf die Analyse kurzer Sequenzen ausgelegt (Ginsburg & Karpiuk,

1994; Oomens et al., 2021; Towse & Neil, 1998). Möchte man dieses Maß für Zufälligkeit für längere Sequenzen berechnen, muss die Komplexität gemittelt werden über die in ggf. überlappende Teilblöcke aufgeteilte Gesamtsequenz. Zenil et al. (2018) haben mittlerweile eine mögliche Erweiterung des Ansatzes vorgestellt, die *Block Decomposition Method* (BDM), die es erlaubt, die algorithmische Komplexität auch für längere Sequenzen zu berechnen. Zwar ist der Begriff der algorithmischen Komplexität eng mit dem Begriff der Zufälligkeit verknüpft, da in einer zufälligen Sequenz algorithmische Regelmäßigkeiten jedweder Art nicht auftreten sollten (Gauvrit et al., 2016). Jedoch fehlt bisher eine systematische Analyse, ob die neu vorgeschlagenen Maße für Komplexität bei der Anwendung auf Sequenzen aus RNG-Aufgaben tatsächlich die Zufälligkeit dieser Sequenzen abzubilden erlauben. Dies herauszuarbeiten ist eines der Ziele der vorliegenden Dissertationsschrift.

1.3 Modellierung von Verhalten beim Generieren von Zufall

In einem kürzlich veröffentlichten Artikel von Yousif et al. (2022) wurde ein mathematisches Modell vorgeschlagen, mit dem man systematische Verhaltensweisen von Menschen bei der Generierung zufälliger Sequenzen als Modellparameter in einer Verhaltensgleichung beschreiben kann. Das Modell wurde in der Publikation von Yousif et al. nur für RNG-Aufgaben definiert, bei denen die Zahlen von 1 bis 9 verwendet werden dürfen. Es ist aber ohne Probleme auf RNG-Aufgaben mit anderen erlaubten Zahlenbereichen erweiterbar. Im Rahmen dieser Dissertationsschrift wird ebenfalls nur der Fall von RNG-Aufgaben behandelt, bei dem die Zahlen von 1 bis 9 erlaubt sind. Das Modell sieht zwei Parameter vor, die systematische Verhaltensweisen ausdrücken, die Menschen beim Generieren von Zufallszahlen zeigen. Der erste Parameter wird im Folgenden als Wiederholungsparameter bezeichnet. Dieser drückt aus, ob eine Person die Tendenz hat, nach Generierung einer Zahl in einer Sequenz diese erneut zu generieren oder deren Wiederholung zu meiden. In der Forschungsliteratur ist es gut dokumentiert, dass Menschen dazu neigen, Wiederholungen beim Generieren von zufälligen Zahlensequenzen zu vermeiden (Cooper, 2016; Ginsburg & Karpiuk, 1994; Peters et al., 2007; Towse, 1998). Der zweite Parameter des Modells nach Yousif et al. wird als Seitenwechselparameter bezeichnet. Dieser drückt aus, ob eine systematische Tendenz beim Generieren von Zufallszahlen dergestalt besteht, dass es in aufeinanderfolgenden Zahlen besonders häufig oder selten zu einem Wechsel von hohen zu niedrigen oder von niedrigen zu hohen Zahlen kommt. In diesem Modell wird die Wahrscheinlichkeit, mit der eine Zahl aus dem Vektor aller *K* möglichen Zahlen in einer Sequenz als nächstes generiert wird, folgendermaßen definiert:

$$\sigma(z)_i = \frac{e^{\epsilon \cdot r_i + \eta \cdot s_i}}{\sum_{j=1}^{K} e^{\epsilon \cdot r_j + \eta \cdot s_j}}. \quad (1)$$

Hierbei ist ϵ der Wiederholungsparameter. Dieser wird mit dem Wert r_i aus dem Vektor r der Länge K multipliziert, der mit dem Wert 1 codiert, ob eine Zahl eine Wiederholung der in der Sequenz zuvor gewählten Zahl ist. Anderenfalls werden die Einträge dieses Vektors mit dem Wert 0 codiert. Dies führt bei einem positiven Wiederholungsparameter dazu, dass eine Wiederholung von Zahlen wahrscheinlicher wird. Negative Parameterwerte hingegen führen dazu, dass direkte Wiederholungen von Zahlen gemieden werden. Parameterwerte von Null drücken aus, dass keine systematische Tendenz zu übermäßig vielen oder wenigen Wiederholungen in einer Sequenz besteht.

Der Seitenwechselparameter wird mit η bezeichnet. Die Seite der niedrigen Zahlen beinhaltet hierbei nach Definition die Zahlen von 1 bis 4, die Seite der hohen Zahlen die Zahlen von 6 bis 9. Die Zahl 5 ist das Mittelelement, welche keine der beiden Seiten zugeordnet ist und somit auch nicht von dem Parameter beeinflusst wird. Der Seitenwechselparameter wird mit dem Vektoreintrag s_i aus dem Vektor s der Länge K multipliziert. Dieser codiert im Eintrag s_i mit dem Wert 1, ob eine Zahl einen Seitenwechsel von einer niedrigen zu einer hohen oder von einer hohen zu einer niedrigen Zahl relativ zur zuvor generierten Zahl in der Sequenz darstellt. Der Wert -1 für einen Vektoreintrag s_i codiert, ob eine Zahl auf derselben Seite ist wie die zuvor generierte Zahl. Der Wert für das Mittelelement 5 im Vektor s ist immer als 0 codiert, da diese Zahl keine der beiden Seiten zuzuordnen ist. Ist die zuvor generierte Zahl einer Sequenz eine 5, sind alle Einträge des Vektor s mit einer 0 codiert, da ein Seitenwechsel nach dem Mittelelement 5 per Definition nicht möglich ist. Positive Werte im Seitenwechselparameter führen folglich zu einer erhöhten Wahrscheinlichkeit, in einer Sequenz in aufeinanderfolgenden Zahlen von einer niedrigen zu einer hohen oder von einer hohen zu einer niedrigen Zahl zu wechseln. Negative Parameterwerte drücken aus, dass eine Tendenz dazu besteht, in aufeinanderfolgenden Zahlen auf

derselben Seite der Zahlen zu verbleiben. Ein Wert von Null in diesem Parameter drückt auch hier aus, dass keine systematische Tendenz zu übermäßig vielen oder wenigen Seitenwechseln besteht.

Die vorgeschlagene Verhaltensgleichung erlaubt es, die oben geschilderten systematischen Verhaltensweisen entsprechend der Logik einer Tendenz zu übermäßig vielen oder wenigen Wiederholungen und/oder Seitenwechseln auszudrücken. Beispielsweise wird bei einem positiven Wiederholungsparameter durch die Formel selektiv die Wahrscheinlichkeit einer direkten Wiederholung erhöht, wodurch die relative Wahrscheinlichkeit aller anderen Zahlen sinkt. Dies wird durch die Definition der Verhaltensgleichung gewährleistet, da das Entscheidungsgewicht, eine Zahl zu wählen (Zähler auf der rechten Seite der Gleichung), geteilt wird durch die Summe der Entscheidungsgewichte aller zur Verfügung stehender Zahlen (Nenner auf der rechten Seite der Gleichung). Dies führt auch dazu, dass die Wahrscheinlichkeiten für alle Zahlen zwischen 0 und 1 liegen und sich zusammen zu 1 aufaddieren.

Der von Yousif et al. (2022) vorgeschlagene Modellierungsansatz lässt sich vielseitig einsetzen. So können die angenommenen latenten Variablen hinter der Zufallszahlengenerierung in einer Verhaltensformel ausgedrückt und direkt als Modellparameter gemessen werden. Außerdem ist es durch diesen Ansatz zur Modellierung möglich, die Passung zwischen beobachteten Zahlensequenzen und dem Modell zu berechnen. Er kann auch dazu verwendet werden, die Passung verschiedener Modelle miteinander zu vergleichen und somit abzuwägen, welches Modell menschengenerierte Zahlensequenzen am besten beschreibt und ob es sinnvoll ist, ein gegebenes Modell mit zusätzlichen Parametern zu erweitern. Dies haben Yousif et al. in einer ersten Validierung des Modells auch unternommen und kamen zu dem Schluss, dass das kombinierte Modell mit dem Wiederholungs- und dem Seitenwechselparameter menschengenerierte Zahlensequenzen aus einer RNG-Aufgabe besser beschreiben konnte als die

Modelle, die jeweils nur einen dieser Parameter beinhalteten. Darüber hinaus haben Yousif et al. das von ihnen definierte Modell dazu benutzt, systematisch Sequenzen von Zahlen zu simulieren, die denen von Menschen ähneln. Durch die Simulation von modellgenerierten Zahlensequenzen ist es möglich, auch Vorhersagen über menschliches Verhalten in RNG-Aufgaben zu machen und diese anschließend empirisch zu überprüfen.

2 Zusammenfassung der Einzelarbeiten

Wie in Kapitel 1 dargestellt, besteht in verschiedenen Gebieten der psychologischen Grundlagenforschung ein Bedarf danach, Zufälligkeit in den von Menschen generierten Sequenzen einer RNG-Aufgabe zu analysieren. Gleichzeitig existiert jedoch eine Vielzahl von Methoden zur Quantifizierung der Zufälligkeit bzw. der Abwesenheit von Zufälligkeit dieser Sequenzen. In der psychologischen Forschung werden häufig die Maße nach Ginsburg und Karpiuk (1994) sowie von Towse und Neil (1998) angewandt (Capone et al., 2014; Cooper, 2016; Maes et al., 2011; Oomens et al., 2021; Peters et al., 2007; Zabelina et al., 2012). In manchen Studien werden jedoch auch alternative Ansätze verfolgt. So wurden nicht nur Ansätze zur Berechnung der Entropie (Barbasz et al., 2008; Jokar & Mikaili, 2012; Oomens et al., 2023), sondern auch Maße algorithmischer Komplexität, wie der kürzliche vorgeschlagene Ansatz zur Approximation der algorithmischen Komplexität für kurze Sequenzen (Biesaga et al., 2021; Biesaga & Nowak, 2022; Gauvrit et al., 2016) sowie auf Kompressionsalgorithmen basierende Methoden angewandt (Wong et al., 2021). Es wurde jedoch wiederholt kritisiert, dass kein Konsens darüber besteht, wie Zufälligkeit in RNG-Aufgaben am besten gemessen werden soll (Barbasz et al., 2008; Gauvrit et al., 2016; Wagenaar, 1972). Das Ziel dieser Dissertationsschrift ist es, einen umfassenden Vergleich zwischen den verschiedenen zur Verfügung stehenden Ansätzen zur Quantifizierung von Zufälligkeit vorzunehmen und herauszuarbeiten, welche dieser Methoden gut und welche weniger gut geeignet sind, um menschliches Verhalten in RNG-Aufgaben hinsichtlich darin bestehender Systematiken zu analysieren. Dabei werden Empfehlungen abgeleitet, die Forschenden bei der Entscheidung für Methoden zur Analyse von menschengenerierten Sequenzen aus RNG-Aufgaben für ihre jeweilige Forschungsfrage helfen sollen.

Die Vielzahl der schon länger zur Verfügung stehenden Methoden zur Messung von Zufälligkeit ergänzt ein jüngst vorgestellter formaler Ansatz zur Modellierung systematischen Verhaltens von Menschen in RNG-Aufgaben von Yousif et al. (2022). Das zweite Ziel dieser Dissertationsschrift ist die Prüfung und Erweiterung dieses Modells zum Zwecke einer Optimierung der Beschreibung und Vorhersage menschengenerierter Zufallszahlen.

2.1 Studie 1: Vergleich von Maßen für Zufälligkeit zur Analyse menschengenerierter Sequenzen

Gegenstand der ersten Studie ist ein umfassender Vergleich verschiedener Maße für Zufälligkeit hinsichtlich ihrer Sensitivität bei der Erkennung von Mustern und Systematiken in menschengenerierten Zahlensequenzen. Dies ist für Forschende insofern eine wichtige Information, als derzeit viele Methoden zur Quantifizierung von Zufälligkeit zur Verfügung stehen und es keinen klaren Konsens darüber gibt, welche dieser Methoden am besten dazu geeignet sind, menschliche Zahlensequenzen aus einer RNG-Aufgabe zu analysieren. Dazu werden in der vorliegenden Forschungsarbeit nicht nur häufig in der Psychologie angewandte Maße für Zufälligkeit untersucht (Towse & Neil, 1998), sondern auch Maße für Entropie aus der Informationstheorie (Moore et al., 2018; Shannon, 1948) sowie Maße für algorithmische Komplexität (Gauvrit et al., 2016; Lempel & Ziv, 1976; Zenil et al., 2018). Ein systematischer Vergleich einer derart großen Anzahl und Vielfalt von Methoden zur Analyse von menschengenerierten Zufallssequenzen wurde bislang noch nicht durchgeführt. Darüber hinaus wird in dieser Dissertationsschrift auch erstmals die für die Anwendung in Forschung und Praxis wichtige Frage untersucht, ob und in welchem Maße die Nützlichkeit der verfügbaren Maße für die Detektion menschentypischer Verhaltensweisen von der Länge der zu untersuchenden Sequenzen abhängt.

Um die Sensitivität der untersuchten Methoden zur Quantifizierung von Zufälligkeit gegenüber menschentypischen Verhaltensweisen zu untersuchen, wurde in dieser Arbeit ein klassifikationsbasierter Ansatz verfolgt. Ob ein Maß sensitiv gegenüber menschentypischen Verhaltensweisen in einer RNG-Aufgabe ist, wurde über die Wahrscheinlichkeit definiert, mit der es möglich ist, zwischen echt zufälligen und menschengenerierten Sequenzen zu differenzieren.

Zu diesem Zweck wurden Sequenzen von Zahlen der Länge 200 im diskreten Intervall von 1 bis 9 aus zwei Quellen generiert:

1) 830 Sequenzen von menschlichen Teilnehmenden in einer Online-Studie, wobei die Zahlen durch Mausklicks auf eine 3x3-Feldertafel generiert wurden, in der die Zahlen in aufsteigender Reihenfolge erst von nach links nach rechts und dann von oben nach unten sortiert waren.

2) Dieselbe Anzahl an zufälligen Sequenzen (= 830) mit derselben Länge (= 200) basierend auf atmosphärischem Rauschen (Eddelbuettel, 2017; Haahr, 2023). Der Vorteil der Verwendung von Zahlensequenzen aus einer annehmbar vollkommen zufälligen Quelle (Furutsu & Ishida, 1961) ist im Vergleich zu computergenerierten pseudorandomisierten Zahlensequenzen, dass auf atmosphärischem Rauschen beruhende Zahlensequenzen nicht deterministisch sind und ihnen keine Periodizität zugrunde liegt (Haahr, 2023).

Ob ein Maß für Zufälligkeit sensitiv gegenüber menschlichen Verhaltensweisen in einer RNG-Aufgabe ist, wurde durch die Anwendung logistischer Regressionsmodelle bestimmt, wobei die unabhängige Variable dieser Analyse das jeweilige Zufälligkeitsmaß war. Die abhängige Variable war die Quelle der Sequenz (menschlich oder zufällig). Um allzu optimistischen Schätzungen der Wahrscheinlichkeit, mit der zwischen menschengenerierten und zufälligen Sequenzen unterschieden werden kann, vorzubeugen, wurde in dieser Forschungsarbeit eine Methode verwendet, die im Englischen als *Bootstrapping* bezeichnet wird (Gine & Zinn, 1990). Dabei wurde das logistische Regressionsmodell zur Unterscheidung der menschengenerierten und zufälligen Sequenzen anhand der einzelnen Zufälligkeitsmaße für eine zufällig gezogene Teilmenge aller vorhanden Sequenzen berechnet. Genauer gesagt wurden zufällig aus der Menge aller Sequenzen so viele Sequenzen gezogen, wie es insgesamt Sequenzen gab. Dabei war es

möglich, dass eine Sequenz mehrmals gezogen wurde und manche Sequenzen gar nicht, da eine gerade gezogene Sequenz sinnbildlich für die Ziehung der nächsten Sequenz in die Menge aller zur Verfügung stehender Sequenzen zurückgelegt wurde. Die Vorhersagekraft des so berechneten Modells wurde dann anhand der Sequenzen evaluiert, die nicht zur Berechnung des Modells verwendet wurden. Diese Prozedur wurde insgesamt 1000 Mal wiederholt; dies erlaubte es zu untersuchen, wie gut ein Maß für Zufälligkeit im Mittel dazu geeignet war, zwischen menschengenerierten und zufälligen Sequenzen zu unterscheiden. Dieses Vorgehen erlaubte außerdem die Konstruktion von Konfidenzintervallen, um ein Maß für die Sicherheit der geschätzten Klassifikationsrate zu erhalten.

Die Ergebnisse dieser ersten Untersuchung erlaubten Schlussfolgerungen darüber, welche Maße für Zufälligkeit gut und welche weniger gut für die Analyse menschengenerierter Sequenzen aus RNG-Aufgaben geeignet sind. Dabei zeigte sich, dass es mehrere Maße für Zufälligkeit gibt, die mit einer Wahrscheinlichkeit von 80% und teilweise über 90% zwischen menschengenerierten und zufälligen Sequenzen unterscheiden konnten. So konnten Maße für algorithmische Komplexität für kurze Sequenzen nach Gauvrit et al. (2016), Maße für Blockentropie sowie Methoden zur Quantifizierung von Wiederholungstendenzen in einer Sequenz am zuverlässigsten Aufschluss darüber geben, ob eine Sequenz von einem Menschen oder einem zufälligen Prozess generiert wurde. Insbesondere Maße für algorithmische Komplexität sowie Methoden zur Quantifizierung von Wiederholungstendenzen über Intervalle von drei bis fünf Zahlen erlaubten eine nahezu perfekte Klassifikation in menschengenerierte und zufällige Sequenzen. Dies passt zu bisherigen Befunden, die eklatante Mängel in der Qualität menschengenerierter Sequenzen in RNG-Aufgaben aufdeckten (Ginsburg & Karpiuk, 1994; Peters et al., 2007; Towse, 1998; Towse & Neil, 1998). An dieser Stelle sei jedoch darauf hingewiesen, dass menschengenerierte

Sequenzen systematisch höhere Werte für algorithmische Komplexität nach Gauvrit et al. (2016) als zufällige Sequenzen erhielten. Dies deutet darauf hin, dass das neu vorgeschlagene Komplexitätsmaß invers mit Zufälligkeit zusammenhängt, was bei der Interpretation dieses Maßes berücksichtigt werden muss. Eine mögliche Erklärung für diesen Befund ist, dass menschentypische Verhaltensweisen wie der Zirkulationstendenz zu einer gleichmäßigen Nutzung aller möglichen Zahlen in bereits kurzen Sequenzabschnitten beiträgt, was zu höheren Werten algorithmischer Komplexität führt. Für eine detaillierte Beschreibung und Interpretation des Ergebnismusters wird auf die erste Einzelarbeit im Anhang verwiesen.

Die gewählte Untersuchungsmethode erlaubte auch erstmals eine systematische Analyse hinsichtlich des Effekts der Sequenzlänge auf die Nützlichkeit der einzelnen Zufälligkeitsmaße bei der Unterscheidung zwischen menschengenerierten und zufälligen Sequenzen. So zeigte sich, dass Maße für algorithmische Komplexität unabhängig von der untersuchten Sequenzlänge gut dazu geeignet sind, zwischen menschengenerierten und zufälligen Sequenzen zu differenzieren. Dies unterstreicht die Nützlichkeit des von Gauvrit et al. (2016) vorgeschlagenen Maßes für algorithmische Komplexität. Bei der Interpretation dieses Maßes muss jedoch berücksichtigt werden, dass besonders hohe Werte in der algorithmischen Komplexität charakteristisch für menschengenerierte und nicht für zufällige Sequenzen zu sein scheinen. Maße für Blockentropie zeigten diese wünschenswerte Eigenschaft nicht: Kurze Sequenzen der Länge 20 erlaubten es nicht, mit überzufällig hoher Wahrscheinlichkeit zwischen menschengenerierten und zufälligen Sequenzen zu differenzieren. Erst bei Betrachtung länger Teilsequenzen mit mindestens 100 Zahlen war es möglich, mit hoher Wahrscheinlichkeit zwischen menschengenerierten und zufälligen Sequenzen zu unterscheiden. Dies ist insofern relevant für die praktische Anwendung, als bisherige Studien, die mit diesem Maß verwandte Methoden zur Quantifizierung von

Zufälligkeit einsetzten, häufig nur die Generierung von Sequenzen der Länge 100 in RNG-Aufgaben erforderten (Friedman & Miyake, 2004; Ginsburg & Karpiuk, 1994; Maes et al., 2011; Miyake et al., 2000; Peters et al., 2007; Towse, 1998; Zabelina et al., 2012). Daraus lässt sich ableiten, dass für die Verwendung bestimmter Maße für Zufälligkeit, insbesondere solcher, die aus Maßen der Entropie abgeleitet wurden, Sequenzen vonnöten sind, die aus mehr als 100 Zeichen bestehen, um menschentypische Verhaltensmuster systematisch zu erkennen.

Neben Erkenntnissen darüber, welche Maße für Zufälligkeit besonders gut für die Analyse menschengenerierter Sequenzen geeignet sind, offenbarte die erste Studie auch, welche Maße weniger gut für die Analyse menschengenerierter Sequenzen geeignet sind. So konnten Zufälligkeitsmaße, die die Anzahl von Wechseln zwischen aufsteigenden und absteigenden Zahlenabfolgen innerhalb einer Sequenz oder die Länge aufsteigender Zahlenfolgen messen, unabhängig von der Sequenzlänge, nur mit einer Wahrscheinlichkeit von etwa 50% zwischen menschengenerierten und zufälligen Sequenzen differenzieren. Dies entspricht der erwarteten Klassifikationsleistung, wenn man bei jeder Sequenz lediglich raten würde, ob diese von einem Menschen oder einem zufälligen Prozess generiert wurde. Diese Maße gehören der Sammlung von Towse und Neil (1998) an und finden häufig Anwendung in der psychologischen Forschung. Die hier vorgelegten Ergebnisse legen die Schlussfolgerung nahe, dass nicht alle der häufig verwendeten Maße für Zufälligkeit tatsächlich ihren Zweck hinsichtlich der Charakterisierung menschlichen Verhaltens hinreichend erfüllen. Die Ergebnisse der erste Studie mahnen daher zur Vorsicht gegenüber oft verwendeten Maßen für Zufälligkeit an, da diese für die Analyse von Sequenzen aus RNG-Aufgaben nicht alle gleich geeignet zu sein scheinen.

Die vorgelegte Studie zeigt auch, dass nicht alle Maße für algorithmische Komplexität gut geeignet sind zur Analyse von Sequenzen aus RNG-Aufgaben. So stellte sich heraus, dass

Kompressionsalgorithmen zwar systematisch zwischen menschengenerierten und zufälligen Sequenzen zu differenzieren vermögen. Jedoch ist diese Diskriminationsleistung für alle untersuchten Sequenzlängen deutlich geringer als die von Maßen für algorithmische Komplexität nach Gauvrit et al. (2016) unter Berücksichtigung der oben diskutierten Interpretation dieses Maßes. Ebenso zeigte sich, dass die Klassifikationsleistung eines erweiterten Maßes für algorithmische Komplexität nach Zenil et al. (2018) stark von verschiedenen Parametern der Analyse wie der Sequenzlänge und der Länge der Blöcke, in die die Sequenz zu Analysezwecken unterteilt wird, abhängig ist.

2.2 Studie 2: Ein erweiterter Ansatz zur Modellierung menschlichen Verhaltens bei der Zufallszahlengenerierung

Das Ziel der zweiten im Rahmen dieser Dissertationsschrift vorgelegten Studie war es, den von Yousif et al. (2022) neu vorgeschlagenen Ansatz zur Modellierung von systematischen Verhaltensweisen von Menschen in RNG-Aufgaben zu prüfen, zu modifizieren und zu erweitern. So werden in dieser Forschungsarbeit zwei Anpassungen des bestehenden Modells in Gleichung 1 vorgeschlagen: 1) die Ersetzung des Seitenwechselparameters durch einen weniger formatabhängigen Distanzparameter und 2) die Erweiterung des Modells um einen Zirkulationsparameter. Im Folgenden werden diese Modifikationen kurz beschrieben.

Zunächst wurde der Seitenwechselparameter durch einen Distanzparameter ersetzt. Der Distanzparameters ermöglicht die Modellierung der Tendenz, in aufeinanderfolgenden Zahlen bevorzugt naheliegende (z. B. 9-8) oder weiter entfernte Zahlen (z. B. 2-9) zu wählen. Dies ist verwandt zur Idee des Seitenwechselparameters, da man einen Seitenwechsel von einer niedrigen zu einer hohen Zahl oder umgekehrt annähernd als eine größere Distanz und das Ausharren auf einer Seite als eine geringere Distanz beschreiben kann. Ein Distanzparameter hat jedoch im Vergleich zu dem Seitenwechselparameter den Vorteil, dass die Abstände zwischen Zahlen genauer differenziert werden können. Mit dem Seitenwechselparameter gehen Informationen über den genauen Abstand zwischen aufeinanderfolgenden Zahlen verloren, da künstlich eine Dichotomisierung in die zwei Seiten von niedrigen und hohen Zahlen vorgenommen wird. Der Distanzparameter hat darüber hinaus auch den Vorteil, dass er weniger abhängig ist von Formateffekten der RNG-Aufgabe, da es auch Studien gibt, in denen die Zahlen einer RNG-Aufgabe in einer 3x3-Matrix zur Auswahl präsentiert werden (Kee et al., 2013; Maes et al., 2011). Hier kann beispielsweise die euklidische Distanz verwendet werden, um den räumlichen Abstand

zwischen zwei Zahlen zu quantifizieren. Ersetzt man den Seitenwechselparameter durch den Distanzparameter, erhält man das folgende Modell:

$$\sigma(z)_i = \frac{e^{\epsilon \cdot r_i + \delta \cdot d_i}}{\sum_{j=1}^{K} e^{\epsilon \cdot r_j + \delta \cdot d_j}}. \quad (2)$$

Hier ist δ der Distanzparameter, der mit der Distanz d_i im Distanzvektor d der Länge K multipliziert wird. Hier und im Folgenden wird für den Vektor d die euklidische Distanz zwischen der zuvor gewählten und allen möglichen Zahlen im 3x3-Antwortformat der hier verwendeten RNG-Aufgabe berechnet. Daraus folgt in diesem Anwendungsfall beispielsweise, dass die Zahl 1 (oben links) näher an der 4 (mittig links) liegt als an der 3 (oben rechts). Bei einem positiven Distanzparameter erhöht sich die Wahrscheinlichkeit, eine Zahl zu wählen, umso mehr, je höher der Abstand der betrachteten Zahl zu der zuvor gewählten Zahl in einer Sequenz ist. Bei einem negativen Distanzparameter erhöht sich die Wahrscheinlichkeit einer Zahl umso mehr, je geringer die Distanz zu der zuvor gewählten Zahl. Bei einem Distanzparameter von 0 besteht kein systematischer Trend hin zu aufeinanderfolgenden Zahlen mit einer besonders hohen oder niedrigen Distanz.

Die zweite Modifikation des Modells nach Yousif et al. (2022) besteht in der Erweiterung des Modells um einen dritten Parameter, der sich einer weiteren Systematik annimmt, die Menschen beim Generieren von zufälligen Sequenzen zeigen, nämlich der Zirkulationstendenz. Wie bereits in Kapitel 1 beschrieben, neigen Menschen dazu, alle verfügbaren Zahlen in einer Sequenz sukzessive abzuarbeiten (Ginsburg & Karpiuk, 1994; Peters et al., 2007). Das bedeutet, dass die Generierung einer Zahl als nächstes Element einer Sequenz umso wahrscheinlicher wird, je länger diese nicht generiert wurde. Dies geht auch aus den Ergebnissen der ersten Studie hervor (siehe Anhang 1). Der Vorteil eines Zirkulationsparameters im Vergleich zu den beiden anderen Parametern des Modells ist, dass mit diesem Parameter nicht nur solche Verhaltensweisen

abgebildet werden können, die aus der unmittelbar vorangegangen Entscheidung für eine Zahl resultieren. Die Modellierung von Zirkulationstendenzen erlaubt vielmehr die Berücksichtigung der gesamten vorangegangen Verhaltenshistorie einer Person in der RNG-Aufgabe. Die Wahrscheinlichkeit eine Zahl im Vektor aller möglichen Zahlen zu wählen, wird in dem erweiterten Modell wie folgt dargestellt:

$$\sigma(z)_i = \frac{e^{\epsilon \cdot r_i + \delta \cdot d_i + \beta \cdot g_i}}{\sum_{j=1}^{K} e^{\epsilon \cdot r_j + \delta \cdot d_j + \beta \cdot g_j}}. \quad (3)$$

Hierbei ist β der Zirkulationsparameter und g der Vektor der Länge K, mit dem für jede mögliche Zahl codiert wird, wie lange diese nicht mehr in der Sequenz verwendet wurde. Wird eine Zahl generiert, wird der jeweilige Abstand g_i auf den Wert 1 gesetzt und der Abstand aller anderen Zahlen wird um eine Einheit erhöht. Der Zirkulationsparameter wird mit dem Abstand g_i, multipliziert; dies führt dazu, dass bei einem positiven Zirkulationsparameter solche Zahlen eine besonders hohe Wahrscheinlichkeit haben, die lange in einer Sequenz nicht mehr verwendet wurden. Ist der Zirkulationsparameter negativ, würde dies bedeuten, dass die Wahrscheinlichkeit, eine Zahl in einer Sequenz zu generieren, mit zunehmendem Abstand sinkt. Dies würde eine Tendenz ausdrücken, weiter zurückliegende Zahlen künftig gar nicht mehr zu berücksichtigen. Ein Parameterwert von 0 steht für das völlige Fehlen einer systematischen Zirkulationstendenz.

Problematisch an dieser Variante des Modells ist, dass die Steigerung des Abstandes zur Verwendung einer letzten Zahl von 2 auf 4 genauso gewichtet wird wie eine Steigerung des Abstandes von 12 auf 14. Es erscheint jedoch plausibel, dass mit immer weiter zunehmendem Abstand der Effekt einer weiteren Erhöhung des Abstandes zunehmend kleiner wird. Dies wird ausgedrückt in dem modifizierten Modell:

$$\sigma(z)_i = \frac{e^{\epsilon \cdot r_i + \delta \cdot d_i + \beta \cdot \log_2 g_i}}{\sum_{j=1}^{K} e^{\epsilon \cdot r_j + \delta \cdot d_j + \beta \cdot \log_2 g_j}}. \quad (4)$$

In diesem Fall wird der Abstand g_i zur letztmaligen Verwendung einer Zahl logarithmisch skaliert, was dazu führt, dass mit zunehmendem Abstand eine Asymptote im Effekt des Zirkulationsparameters auf die Entscheidung für eine Zahl erreicht wird. So kann ausgedrückt werden, dass die Erhöhung eines bereits langen Abstandes sich weniger auf die Wahrscheinlichkeit der darauffolgenden Zahl auswirkt als die Erhöhung des Abstandes einer Zahl, deren letzte Nennung weniger lang zurückliegt.

Um die vorgeschlagenen Modelle hinsichtlich ihrer Passung auf menschliches Verhalten bei der Generierung von Zufallszahlen zu überprüfen, wurde der Datensatz aus der ersten Studie wiederverwendet. Die Ergebnisse der Modellvergleiche zeigten, dass die hier vorgeschlagenen Modifikationen des Modells nach Yousif et al. (2022) eine bessere Beschreibung menschlichen Verhaltens in einer RNG-Aufgabe ermöglichten. So zeigte sich in den Ergebnissen, dass das modifizierte Modell mit einem Distanzparameter anstelle des Seitenwechselparameters mit einer höheren Wahrscheinlichkeit erlaubte vorherzusagen, welche Zahl von einer Person als nächstes in einer Sequenz gewählt wird. Weitere Analysen zeigten, dass sich die Vorhersagekraft noch deutlich weiter verbessern ließ, wenn ein Zirkulationsparameter zu dem modifizierten Modell mit dem Distanzparameter hinzugefügt wurde. Des Weiteren zeigte sich, dass eine Skalierung des Abstandes zur letzten Generierung einer Zahl mit dem Logarithmus für den Zirkulationsparameter eine bessere Annäherung an das menschliche Verhalten erlaubte als eine lineare Skalierung. Die Ergebnisse bestätigten, dass Menschen dazu tendieren, zu schnell durch alle verfügbaren Zahlen einer Sequenz zu zirkulieren. Dabei steigerte die Erhöhung eines bereits langen Abstandes zur letzten Generierung einer Zahl die Wahrscheinlichkeit der erneuten Wahl dieser Zahl weniger als die Erhöhung des Abstandes einer Zahl, deren letzte Nennung weniger lang zurücklag. Darüber

hinaus zeigten die Ergebnisse der Studie, dass Menschen einerseits eine Tendenz dazu haben, Wiederholungen zu meiden, und andererseits Paare einander naheliegender Zahlen zu generieren. Dies steht im Einklang mit vorheriger Forschung, die solche Verhaltenstendenzen feststellen konnte (Ginsburg & Karpiuk, 1994; Towse, 1998).

In einem nächsten Untersuchungsschritt wurden die Ergebnisse dieser Studie durch modellgetriebene Computerstimulationen erweitert. Ziel der Computersimulationen war es, die Ähnlichkeit menschengenerierter und modellgenerierter Sequenzen hinsichtlich systematischer Muster zu untersuchen. Dieser Ansatz erlaubte es, die Vorhersagen der untersuchten Modellvarianten systematisch zu testen. Die Idee für diesen Ansatz basiert auf der Arbeit von Yousif et al. (2022). Dazu wurden mit den hier untersuchten Modellen Zahlensequenzen simuliert, für deren Grundlage dieselben Parameterwerte verwendet wurden, die für die menschengenerierten Zahlensequenzen geschätzt wurden. Zur Charakterisierung der mit den Modellen simulierten Sequenzen wurden solche Maße für Zufälligkeit verwendet, die sich in der ersten Forschungsarbeit als besonders sensitiv gegenüber menschentypischen Verhaltensweisen erwiesen hatten. So zeigten die mit dem erweiterten Modell mit dem Wiederholungs-, Distanz- und Zirkulationsparameter generierten Sequenzen eine vergleichbare mangelnde Häufigkeit von Wiederholungen über moderat lange Intervalle von Zahlen auf. Dabei war diese Annäherung deutlich besser als durch das Modell mit dem Wiederholungs- und dem Distanzparameter ohne den Zirkulationsparameter. Die Vielseitigkeit des hier verwendeten Modellierungsansatzes zeigt sich darin, dass dieser nicht nur zum Beschreiben beobachteter Zahlensequenzen genutzt werden kann, sondern sich auch als generatives Modell eignet, um Zahlensequenzen zu produzieren, denen systematische Muster zugrunde liegen, die Menschen typischerweise zeigen.

In einer abschließenden Computersimulation wurden systematisch mit dem Modell Sequenzen mit zufällig festgelegten Parameterwerten simuliert. Anschließend wurde untersucht, ob die anhand dieser simulierten Sequenzen berechneten Parameterwerte den zur Simulation verwendeten Parameterwerten entsprechen. Die Ergebnisse zeigten klar, dass alle hier untersuchten Parameter präzise und unabhängig voneinander geschätzt werden können, was die Verlässlichkeit der in dieser Forschungsarbeit geschätzten Parameterwerte unterstreicht. Zusammengenommen untermauern diese Befunde die Nützlichkeit des hier untersuchten Modellierungsansatzes zur Beschreibung und Vorhersage menschlichen Verhaltens.

3 Diskussion

Ein bereits wiederholt angebrachter Kritikpunkt an der Erforschung von Zufälligkeit in menschengenerierten Sequenzen betrifft das hohe Maß an Heterogenität und die mangelnde Einheitlichkeit der verwendeten Methoden zur Quantifizierung von Zufälligkeit (Barbasz et al., 2008; Gauvrit et al., 2016; Wagenaar, 1972). Um diesem Problem zu begegnen, wurden in dieser Dissertationsschrift in einem ersten Schritt in einer umfangreichen Studie verschiedene Methoden zur Quantifizierung von Zufälligkeit in menschengenerierten Zahlensequenzen aus RNG-Aufgaben untersucht und miteinander verglichen. Dabei wurde als Validierungskriterium zugrunde gelegt, wie gut ein Maß für Zufälligkeit dazu in der Lage ist, zwischen menschengenerierten und zufälligen Sequenzen zu unterscheiden. In einem zweiten Schritt wurde ein neuer Ansatz zur formalen Modellierung von systematischen Verhaltensweisen in RNG-Aufgaben modifiziert und erweitert.

In der hier vorgestellten Dissertationsschrift wurde die bisher umfangreichste Analyse von verschiedenen Maßen zur Quantifizierung von Zufälligkeit durchgeführt. So wurden nicht nur die häufig in der psychologischen Forschung verwendeten Maße von Towse und Neil (1998) untersucht, sondern auch eine Vielzahl von Methoden, die ihren Ursprung in der Informatik und Mathematik haben. Zu nennen sind hier Methoden aus der Informationstheorie (Shannon, 1948) sowie Methoden zur Approximation der algorithmischen Komplexität von Sequenzen (Gauvrit et al., 2016; Lempel & Ziv, 1976; Zenil et al., 2018). Zum Vergleich dieser verschiedenen Maße wurde in dieser Arbeit ein klassifikationsbasierter Ansatz verwendet, der es erlaubt, zu untersuchen, wie gut es anhand eines Maßes für Zufälligkeit möglich ist, zwischen menschengenerierten und zufälligen Sequenzen zu unterscheiden. Dabei verließen wir uns nicht auf computergenerierte pseudorandomisierte Zahlensequenzen als Vergleichsmaßstab, sondern

verwendeten zufällige Sequenzen auf Basis atmosphärischen Rauschens (Haahr, 2023). Somit war es uns möglich, einen Goldstandard für zufällige Sequenzen zu etablieren, was dem hier durchgeführten Vergleich ein höheres Gewicht verleiht. Darüber hinaus ist die vorliegende Arbeit nicht nur auf die Analyse der gesamten Sequenzen beschränkt, sondern es wurde auch systematisch untersucht, inwiefern die Nützlichkeit eines Maßes, zwischen menschengenerierten und zufälligen Zahlensequenzen zu unterscheiden, von der Länge der zur Analyse verwendeten Sequenzen beeinflusst wird.

In den durchgeführten Analysen der ersten Forschungsarbeit (siehe Anhang 1) zeigte sich, dass insbesondere Maße algorithmischer Komplexität für kurze Sequenzen nach Gauvrit et al. (2016) gut dazu geeignet waren, menschengenerierte von zufälligen Sequenzen zu unterscheiden. Dieses Maß erlaubte eine nahezu perfekte Differenzierung, und das bereits bei kurzen Sequenzen von nur 20 Zahlen. Jedoch konnten auch einige weitere Maße aus der in der psychologischen Forschung oft angewandten Sammlung von Maßen nach Towse und Neil (1998) eine hohe Klassifikationsleistung erreichen, die ebenfalls auch schon bei relativ kurzen Sequenzen hoch war. Zu nennen sind hier Maße für die Anzahl der Wiederholungen von Zahlen in einer Sequenz über Abstände von etwa drei bis fünf Zahlen hinweg sowie Maße zur Messung der Zirkulationsgeschwindigkeit durch alle Zahlen oder den mittleren Abstand zwischen identischen Zahlen in einer Sequenz. Auch Maße für Block-Entropie aus der Informationstheorie zeigten insbesondere für kurze bis moderat lange Blöcke eine hohe Klassifikationsleistung zwischen menschengenerierten und zufälligen Sequenzen. Es ist jedoch hervorzuheben, dass bei letzterem die Klassifikationsleistung erst bei Betrachtung langer Sequenzen (länger als 100 Zahlen) hoch war. Dieser Befund ist insofern für die gängige Forschungspraxis relevant, da viele Studien oft nur die Generierung von Zahlensequenzen der Länge 100 in RNG-Aufgaben erfordern (Ginsburg &

Karpiuk, 1994; Maes et al., 2011; Peters et al., 2007; Towse, 1998). Hier wäre eine Erhöhung der Sequenzlänge sinnvoll, um das volle Potential der Maße auszuschöpfen, die auf Entropie und deren Varianten basieren. Einige weitere Maße, beispielweise zur Quantifizierung der Wechsel zwischen auf- und absteigenden Zahlensequenzen oder Kompressionsalgorithmen, stellten sich in dieser Untersuchung als weniger nützlich heraus, um zwischen menschengenerierten und zufälligen Sequenzen zu differenzieren.

Die hier vorgestellten Untersuchungsergebnisse sollen dabei helfen, Forschende, die planen, RNG-Aufgaben in ihrer Forschung durchzuführen, über die Eigenschaften der hierfür gebräuchlichen Maße für Zufälligkeit zu informieren. Die Ergebnisse aus der ersten Studie legen nahe, dass es einige Zufälligkeitsmaße gibt, die tatsächlich hoch sensitiv gegenüber systematischem Verhalten von Menschen in RNG-Aufgaben sind. Allerdings sind auch manche der Maße für Zufälligkeit, die häufig in dieser Forschung Anwendung finden, entweder generell nicht sensitiv genug oder in ihrer Leistung stark von der Länge der generierten Sequenzen abhängig. Die hier vorgestellten Ergebnisse sollen dabei helfen, sich in der Vielzahl der zur Verfügung stehenden Maße für Zufälligkeit zurechtzufinden und solche Maße zu identifizieren, die besonders sensitiv gegenüber menschlichem Verhalten sind.

Aus den Ergebnissen lässt sich erkennen, dass insbesondere hinsichtlich der Maße für algorithmische Komplexität noch weitergehender Forschungsbedarf besteht. So zeigten die Ergebnisse dieser Untersuchung, dass menschengenerierte Sequenzen oft eine höhere Komplexität gemittelt über kurze Sequenzblöcke hinweg (Gauvrit et al., 2016) erreichten als tatsächlich zufällige Sequenzen. Dies stellt nicht in Abrede, dass diese Maße eine hohe Diskriminationsleistung zwischen menschengenerierten und zufälligen Sequenzen erzielten. Jedoch muss dieser Befund bei der Interpretation dieses Maßes berücksichtigt werden. Des

Weiteren muss an dieser Stelle auch angemerkt werden, dass andere Maße für algorithmische Komplexität wie Kompressionsalgorithmen oder die Verallgemeinerung der algorithmischen Komplexität nach Gauvrit et al. (2016) für längere Sequenzen (BDM; Zenil et al., 2018) weniger gute und konsistente Ergebnisse bei der Differenzierung zwischen menschengenerierten und zufälligen Sequenzen erzielten. Von einer Verwendung dieser Maße in zukünftigen Studien ist daher abzuraten.

In der zweiten hier präsentierten Forschungsarbeit wurde ein komplementärer Ansatz zur Untersuchung von menschengenerierten Sequenzen in RNG-Aufgaben verbessert. Hierbei wurden mathematische Modelle zur Schätzung der zugrundliegenden systematischen Fehler eingesetzt, die Menschen bei der Generierung von zufälligen Zahlensequenzen machen. Dazu wurde das von Yousif et al. (2022) vorgeschlagene Modell modifiziert und durch einen weiteren Parameter erweitert. Genauer gesagt wurde der Seitenwechselparameter im Ursprungsmodell durch einen Distanzparameter ersetzt, der weniger vom Format der RNG-Aufgabe abhängig ist. Das mit dem Distanzparameter erweiterte Modell vermochte systematische Muster in den menschengenerierten Sequenzen besser zu erklären als das Ursprungsmodell. Außerdem wurde dem Modell ein Zirkulationsparameter hinzugefügt, der es erlaubt, einen unter Menschen stark ausgeprägten Fehler bei der Generierung von Zufallszahlen darzustellen. So neigen Menschen dazu, zu schnell durch alle möglichen Zahlen in einer Sequenz zu zirkulieren, was dazu führt, dass Zahlen, die länger nicht mehr generiert wurden, mit einer erhöhten Wahrscheinlichkeit als nächstes generiert werden (Ginsburg & Karpiuk, 1994). Im Gegensatz zu den bisherigen Parametern des Modells wird hiermit eine Entscheidung nun mehr nicht nur als Reaktion auf die zuvor generierte Zahl dargestellt, sondern als Resultat der gesamten zuvor generierten Zahlensequenz. Die hier

vorgestellten Ergebnisse deuten klar darauf hin, dass die Erweiterung des Modells um diesen Parameter es erlaubt, menschliches Verhalten in RNG-Aufgaben besser zu beschreiben.

In verschiedenen Studien wurden oft eine Vielzahl von Maßen für Zufälligkeit zur Analyse menschlichen Verhaltens in RNG-Aufgaben verwendet (Ginsburg & Karpiuk, 1994; Maes et al., 2011; Peters et al., 2007; Towse & Neil, 1998). Der hier vorgestellte und erweiterte Modellierungsansatz hat im Vergleich dazu mehrere Vorteile. Zum einen erlaubt der Modellierungsansatz durch dessen Parameter eine direkte und gleichzeitige Bestimmung der Fehler, die Menschen beim Generieren von zufälligen Sequenzen machen. Berechnet man im Vergleich dazu verschiedene Maße für Zufälligkeit, ergibt sich das Problem, dass diese nicht voneinander unabhängig sind. So ist beispielsweise die Anzahl der beobachteten Wiederholungen in einer Sequenz abhängig von der Stärke der Zirkulationstendenz einer Person. Bei einer starken Zirkulationstendenz werden bevorzugt solche Zahlen generiert, die lange nicht mehr verwendet wurden und solche Zahlen gemieden, die kürzlich generiert wurden. Dies führt ebenfalls zu einer Meidung von Wiederholungen, was demonstriert, dass die Anzahl der Wiederholungen in einer Sequenz mit der Zirkulationstendenz konfundiert ist. Der Modellierungsansatz hingegen erlaubt es, zwischen diesen beiden Verhaltenskomponenten zu trennen, was eine reinere Schätzung der zugrundeliegenden Fehler beim Generieren von zufälligen Sequenzen ermöglicht.

Darüber hinaus ist der Modellierungsansatz sparsamer, da die angenommenen latenten Variablen hinter der Zufallszahlengenerierung direkt gemessen werden und nicht erst eine Vielzahl von 10 (Ginsburg & Karpiuk, 1994) oder mehr Zufälligkeitsmaßen (Towse & Neil, 1998) berechnet werden muss, die eine anschließende Aggregation erfordern, um Aussagen über die zugrundeliegenden latenten Variablen treffen zu können (Oomens et al., 2023). Außerdem erlaubt es der Modellierungsansatz zu überprüfen, ob die Erweiterung eines Modells zu einer verbesserten

Beschreibung der beobachteten menschlichen Zahlensequenzen führt. Dieser Ansatz wurde auch in der hier vorgestellten Forschungsarbeit angewandt, um den Distanz- und Zirkulationsparameter in das Modell von Yousif et al. (2022) zu integrieren. Diese Methodik ist insbesondere für die zukünftige Forschung nützlich, um systematisch zu untersuchen, welche Modelle menschlichen Verhaltens die beobachten Systematiken aus RNG-Aufgaben am besten beschreiben können. Zuletzt kann das Modell auch genutzt werden, um Zahlensequenzen zu generieren, die denen von Menschen ähneln. Dies kann dazu verwendet werden, um Modellvorhersagen anhand empirischer Daten zu überprüfen. Die Ergebnisse einer solchen hier durchgeführten Simulation legen nahe, dass das in dieser Dissertationsschrift vorgeschlagene Modell gut dazu genutzt werden kann, systematische Abweichungen vom Zufall in menschengenerierten Zahlensequenzen zu erklären.

An dieser Stelle muss betont werden, dass der vorgeschlagene Modellierungsansatz nicht zwangsläufig in direkter Konkurrenz mit dem in der ersten Forschungsarbeit untersuchten Methoden zur direkten Quantifizierung von Zufälligkeit in Sequenzen steht. So ist es von der jeweiligen Forschungsfrage abhängig, welcher Ansatz am besten zur Untersuchung von Zufälligkeit in menschengenerierten Sequenzen geeignet ist. Für die Untersuchung spezifischer Fehler beim Generieren von zufälligen Sequenzen, wie der Tendenz zu Wiederholungen, der Tendenz zu naheliegenden Wertpaaren oder der Tendenz, zu schnell durch alle möglichen Zahlen zu zirkulieren, ist das hier erweiterte Modell mit dem Wiederholungs-, Distanz-, und Zirkulationsparameter zu empfehlen. Der Vorteil dieses Modells gegenüber der bloßen Berechnung bestimmter Maße für Zufälligkeit besteht darin, dass die zugrundeliegenden Systematiken im Verhalten gleichzeitig und unabhängig voneinander geschätzt werden können. Abhängig vom Kontext kann es jedoch auch wünschenswert sein, für menschengenerierte Sequenzen aus RNG-Aufgaben besonders sensitive Maße zu berechnen, wie beispielsweise Maße

für algorithmische Komplexität nach Gauvrit et al. (2016), um Veränderungen im Verhalten zu messen, die sich nicht nur in spezifischen Systematiken in Zahlensequenzen wie der Anzahl der Wiederholungen zeigen.

Als Limitation dieser Dissertationsschrift muss genannt werden, dass verschiedene Formate von RNG-Aufgaben im Rahmen der vorgelegten Forschungsarbeiten nicht miteinander verglichen wurden. Die Studienteilnehmenden produzierten Zahlensequenzen, indem sie die Zahlen von 1 bis 9 per Mausklick in einer 3x3-Feldertafel auswählten. In einigen anderen Studien wurde jedoch ein mündliches Antwortformat für die Generierung der Zahlensequenzen verwendet (Figurska et al., 2008; Ginsburg & Karpiuk, 1994; Schulz et al., 2021). Es ist möglich, dass die Verwendung einer 3x3-Feldertafel im Vergleich zu einer mündlichen Generierung von Zahlensequenzen zu anderen Verhaltensweisen in einer RNG-Aufgabe führt. Eine denkbare Konsequenz dieses Formatunterschiedes wäre, dass numerisch aufeinanderfolgende Zahlenpaare wie 3-4 oder 4-3 seltener bei dem in dieser Forschungsarbeit verwendetem Antwortformat generiert werden, weil in einer 3x3-Feldertafel die Zahlen 3 und 4 nicht mehr direkt miteinander benachbart sind. So ist in der hier verwendeten 3x3-Feldertafel die 3 rechts oben und die 4 mittig links platziert. Diese Vermutung steht im Einklang mit den Befunden nach Maes et al. (2013). Diese kamen zwar zum Ergebnis, dass beide RNG-Antwortformate zu vergleichbaren Ergebnissen in den Zufälligkeitsmaßen nach Towse und Neil (1998) führten. Jedoch wurde als Einschränkung festgestellt, dass in der Aufgabe, in der die Zahlen per Mausklick generiert wurden, weniger benachbarten Zahlenpaare in den Sequenzen generiert wurden als in einer mündlichen RNG-Aufgabe. Damit lag der prozentuale Anteil von benachbarten Zahlenpaaren in der Mausklick-Aufgabe näher am theoretischen Erwartungswert zufälliger Sequenzen. Eine mögliche Lösung für dieses Problem ist der hier erweiterte Ansatz zur mathematischen Modellierung der

zugrundeliegenden Verhaltensweisen bei RNG-Aufgaben, der es erlaubt, mithilfe des Distanzparameters Formateffekte in aufeinanderfolgenden Zahlen in Sequenzen zu modellieren. Es wäre denkbar, dass sich die Tendenz zu benachbarten Zahlenpaaren in der Mausklick-Aufgabe mit der 3x3-Feldertafel nicht mehr primär in numerisch aufeinanderfolgenden Zahlenpaaren, sondern in der bevorzugten Wahl von räumlich benachbarten Feldern bemerkbar macht. Zukünftige Studien sollten solche Formateffekte mithilfe des hier erweiterten Modellierungsansatzes mit dem Distanzparameter adressieren.

Darüber hinaus wurde in dieser Arbeit bei der Generierung der Zahlensequenzen in der RNG-Aufgabe nur ein Zeichensatz mit den Zahlen von 1 bis 9 betrachtet. Der Grund hierfür war, dass dieses Antwortformat in verschiedenen Studien zur menschlichen RNG-Forschung häufig verwendet wird (Capone et al., 2014; Jokar & Mikaili, 2012; Miyake et al., 2000; Schulz et al., 2012, 2021; Zabelina et al., 2012). Jedoch werden in manchen Studien auch Varianten von RNG-Aufgaben mit nur drei (Wong et al., 2021) oder zwei Antwortalternativen (Biesaga et al., 2021; Bocharov et al., 2020; Gauvrit et al., 2016; Shteingart & Loewenstein, 2016) verwendet. Es wäre denkbar, dass abhängig von der Größe des verwendeten Zeichensatzes verschiedene Methoden zur Messung von Zufälligkeit unterschiedlich sensitiv gegenüber menschentypischen Verhaltensweisen in den generierten Zahlensequenzen sind. Jedoch scheinen die Befunde aus bisherigen Studien mit einem kleineren Zeichensatz im Einklang mit den hier vorgestellten Ergebnissen zu stehen. So kamen Gauvrit et al. und Biesaga et al. ebenfalls zu dem Ergebnis, dass Maße algorithmischer Komplexität hoch sensitiv gegenüber menschentypischen Verhaltensweisen bei binären RNG-Aufgaben sind. Darüber hinaus könnte sich zukünftige Forschung insbesondere auch mit der Fragestellung befassen, wie der Ansatz zur mathematischen Modellierung von

systematischen Verhaltensweisen in RNG-Aufgaben mit einem binären Antwortformat angewandt werden könnte.

Zusammenfassend war es das Ziel dieser Dissertationsschrift, verschiedene Methoden zur Messung und mathematischen Modellierung von Zufälligkeit in menschengenerierten Zahlensequenzen zu untersuchen und zu vergleichen. Diese Arbeit geht dabei insofern über bestehende Forschung hinaus, als die bislang umfassendste Sammlung von Methoden zur Messung von Zufälligkeit aus verschiedenen theoretischen Ansätzen analysiert und miteinander verglichen wurde. Darüber hinaus wurde ein erst kürzlich vorgeschlagener Ansatz zur Modellierung der bei dieser Aufgabe zugrundeliegenden Verhaltensmechanismen modifiziert und so erweitert, dass eine deutlich verbesserte Vorhersage menschengenerierter Zufallszahlen möglich wurde. Die hier vorgestellten Ergebnisse ermöglichen Forschenden künftig eine besser informierte Wahl hinsichtlich der besten zur Verfügung stehenden Methoden zur Untersuchung menschengenerierter Zufallszahlen, sowie eine verbesserte Modellierung der dabei wirksamen Prozesse.

Literaturverzeichnis

Baddeley, A. D. (1966). The capacity for generating information by randomization. *Quarterly Journal of Experimental Psychology*, *18*(2), 119–129. https://doi.org/10.1080/14640746608400019

Barbasz, J., Stettner, Z., Wierzchoń, M., Piotrowski, K. T., & Barbasz, A. (2008). How to estimate the randomness in random sequence generation tasks? *Polish Psychological Bulletin*, *39*(1), 42–46. https://doi.org/10.2478/v10059-008-0006-7

Biesaga, M., & Nowak, A. (2022). *The role of the working memory storage component in the random-likes series generation*. PsyArXiv. https://doi.org/10.31234/osf.io/bvjkw

Biesaga, M., Talaga, S., & Nowak, A. (2021). The effect of context and individual differences in human-generated randomness. *Cognitive Science*, *45*(12), e13072. https://doi.org/10.1111/cogs.13072

Bocharov, A., Freedman, M., Kemp, E., Roetteler, M., & Svore, K. M. (2020). *Predicting human-generated bitstreams using classical and quantum models*. arXiv. https://doi.org/10.48550/ARXIV.2004.04671

Brugger, P., Monsch, A. U., Salmon, D. P., & Butters, N. (1996). Random number generation in dementia of the Alzheimer type: A test of frontal executive functions. *Neuropsychologia*, *34*(2), 97–103. https://doi.org/10.1016/0028-3932(95)00066-6

Capone, F., Capone, G., Ranieri, F., Di Pino, G., Oricchio, G., & Di Lazzaro, V. (2014). The effect of practice on random number generation task: A transcranial direct current stimulation study. *Neurobiology of Learning and Memory*, *114*, 51–57. https://doi.org/10.1016/j.nlm.2014.04.013

Cooper, R. P. (2016). Executive functions and the generation of "random" sequential responses: A computational account. *Journal of Mathematical Psychology*, *73*, 153–168. https://doi.org/10.1016/j.jmp.2016.06.002

Eddelbuettel, D. (2017). *random: True random numbers using RANDOM.ORG* [Computer software]. https://CRAN.R-project.org/package=random

Figurska, M., Stańczyk, M., & Kulesza, K. (2008). Humans cannot consciously generate random numbers sequences: Polemic study. *Medical Hypotheses*, *70*(1), 182–185. https://doi.org/10.1016/j.mehy.2007.06.038

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General*, *133*(1), 101–135. https://doi.org/10.1037/0096-3445.133.1.101

Furutsu, K., & Ishida, T. (1961). On the theory of amplitude distribution of impulsive random noise. *Journal of Applied Physics*, *32*(7), 1206–1221. https://doi.org/10.1063/1.1736206

Gauvrit, N., Singmann, H., Soler-Toscano, F., & Zenil, H. (2016). Algorithmic complexity for psychology: A user-friendly implementation of the coding theorem method. *Behavior Research Methods*, *48*(1), 314–329. https://doi.org/10.3758/s13428-015-0574-3

Gauvrit, N., Zenil, H., Delahaye, J.-P., & Soler-Toscano, F. (2014). Algorithmic complexity for short binary strings applied to psychology: A primer. *Behavior Research Methods*, *46*(3), 732–744. https://doi.org/10.3758/s13428-013-0416-0

Gauvrit, N., Zenil, H., Soler-Toscano, F., Delahaye, J.-P., & Brugger, P. (2017). Human behavioral complexity peaks at age 25. *PLOS Computational Biology*, *13*(4), e1005408. https://doi.org/10.1371/journal.pcbi.1005408

Gine, E., & Zinn, J. (1990). Bootstrapping General Empirical Measures. *The Annals of Probability*, *18*(2), 851–869. http://www.jstor.org/stable/2244320

Ginsburg, N., & Karpiuk, P. (1994). Random generation: Analysis of the responses. *Perceptual and Motor Skills*, *79*(3), 1059–1067. https://doi.org/10.2466/pms.1994.79.3.1059

Haahr, M. (2023). *RANDOM.ORG: True random number service* [Computer software]. https://www.random.org

Heuer, H., Janczyk, M., & Kunde, W. (2010). Random noun generation in younger and older adults. *Quarterly Journal of Experimental Psychology*, *63*(3), 465–478. https://doi.org/10.1080/17470210902974138

Heuer, H., Kohlisch, O., & Klein, W. (2005). The effects of total sleep deprivation on the generation of random sequences of key-presses, numbers and nouns. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *58*(2), 275–307. https://doi.org/10.1080/02724980343000855

Howarth, S., Handley, S. J., & Walsh, C. (2016). The logic-bias effect: The role of effortful processing in the resolution of belief–logic conflict. *Memory and Cognition*, *44*(2), 330–349. https://doi.org/10.3758/s13421-015-0555-x

Jokar, E., & Mikaili, M. (2012). Assessment of human random number generation for biometric verification. *Journal of Medical Signals and Sensors*, *2*(2), 82–87. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3632045/

Kee, Y. H., Chaturvedi, I., Wang, C. K. J., & Chen, L. H. (2013). The power of now: Brief mindfulness induction led to increased randomness of clicking sequence. *Motor Control*, *17*(3), 238–255. https://doi.org/10.1123/mcj.17.3.238

Knott, L. M., & Dewhurst, S. A. (2007). The effects of divided attention at study and test on false recognition: A comparison of DRM and categorized lists. *Memory and Cognition*, *35*(8), 1954–1965. https://doi.org/10.3758/BF03192928

Lempel, A., & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on Information Theory*, *22*(1), 75–81. https://doi.org/10.1109/TIT.1976.1055501

Maes, J. H. R., Eling, P. A. T. M., Reelick, M. F., & Kessels, R. P. C. (2011). Assessing executive functioning: On the validity, reliability, and sensitivity of a click/point random number generation task in healthy adults and patients with cognitive decline. *Journal of Clinical and Experimental Neuropsychology*, *33*(3), 366–378. https://doi.org/10.1080/13803395.2010.524149

Marsh, J. E., Sörqvist, P., Halin, N., Nöstl, A., & Jones, D. M. (2013). Auditory distraction compromises random generation: Falling back into old habits? *Experimental Psychology*, *60*(4), 279–292. https://doi.org/10.1027/1618-3169/a000198

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Miyake, A., Witzki, A. H., & Emerson, M. J. (2001). Field dependence–independence from a working memory perspective: A dual-task investigation of the Hidden Figures Test. *Memory*, *9*(4–6), 445–457. https://doi.org/10.1080/09658210143000029

Moore, D. G., Valentini, G., Walker, S. I., & Levin, M. (2018). Inform: Efficient information-theoretic analysis of collective behaviors. *Frontiers in Robotics and AI*, *5*. https://doi.org/10.3389/frobt.2018.00060

Oomens, W., Maes, J. H. R., Hasselman, F., & Egger, J. I. M. (2021). RandseqR: An R package for describing performance on the random number generation task. *Frontiers in Psychology*, *12*, 629012. https://doi.org/10.3389/fpsyg.2021.629012

Oomens, W., Maes, J. H. R., Hasselman, F., & Egger, J. I. M. (2023). A time-series perspective on executive functioning: The benefits of a dynamic approach to random number generation. *International Journal of Methods in Psychiatric Research*, *32*(2), e1945. https://doi.org/10.1002/mpr.1945

Peters, M., Giesbrecht, T., Jelicic, M., & Merckelbach, H. (2007). The random number generation task: Psychometric properties and normative data of an executive function task in a mixed sample. *Journal of the International Neuropsychological Society*, *13*(4), 626–634. https://doi.org/10.1017/S1355617707070786

Schulz, M.-A., Baier, S., Timmermann, B., Bzdok, D., & Witt, K. (2021). A cognitive fingerprint in human random number generation. *Scientific Reports*, *11*(1), 20217. https://doi.org/10.1038/s41598-021-98315-y

Schulz, M.-A., Schmalbach, B., Brugger, P., & Witt, K. (2012). Analysing humanly generated random number sequences: A pattern-based approach. *PLoS ONE*, *7*(7), e41531. https://doi.org/10.1371/journal.pone.0041531

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shinba, T., Shinozaki, T., Kariya, N., & Ebata, K. (2000). Random number generation deficit in schizophrenia characterized by oral vs written response modes. *Perceptual and Motor Skills*, *91*(4), 1091–1105. https://doi.org/10.2466/pms.2000.91.3f.1091

Shteingart, H., & Loewenstein, Y. (2016). Heterogeneous suppression of sequential effects in random sequence generation, but not in operant learning. *PLOS ONE*, *11*(8), e0157643. https://doi.org/10.1371/journal.pone.0157643

Soler-Toscano, F., Zenil, H., Delahaye, J.-P., & Gauvrit, N. (2014). Calculating Kolmogorov complexity from the output frequency distributions of small Turing machines. *PLOS ONE*, *9*(5), e96223. https://doi.org/10.1371/journal.pone.0096223

Towse, J. N. (1998). On random generation and the central executive of working memory. *British Journal of Psychology*, *89*(1), 77–101. https://doi.org/10.1111/j.2044-8295.1998.tb02674.x

Towse, J. N., & Cheshire, A. (2007). Random number generation and working memory. *European Journal of Cognitive Psychology*, *19*(3), 374–394. https://doi.org/10.1080/09541440600764570

Towse, J. N., & Neil, D. (1998). Analyzing human random generation behavior: A review of methods used and a computer program for describing performance. *Behavior Research Methods, Instruments, & Computers*, *30*(4), 583–591. https://doi.org/10.3758/BF03209475

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110. https://doi.org/10.1037/h0031322

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin*, *77*(1), 65–72. https://doi.org/10.1037/h0032060

Williams, I. A., Obeso, I., & Jahanshahi, M. (2020). Dopaminergic medication improves cognitive control under low cognitive demand in Parkinson's disease. *Neuropsychology*, *34*(5), 551–559. https://doi.org/10.1037/neu0000629

Wong, A., Merholz, G., & Maoz, U. (2021). Characterizing human random-sequence generation in competitive and non-competitive environments using Lempel–Ziv complexity. *Scientific Reports*, *11*(1), 20662. https://doi.org/10.1038/s41598-021-99967-6

Yousif, S. R., McDougle, S. D., & Rutledge, R. B. (2022). A task-general model of human randomization. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*. https://escholarship.org/uc/item/978107w8

Zabelina, D. L., Robinson, M. D., Council, J. R., & Bresin, K. (2012). Patterning and nonpatterning in creative cognition: Insights from performance in a random number generation task. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(2), 137–145. https://doi.org/10.1037/a0025452

Zenil, H., Hernández-Orozco, S., Kiani, N. A., Soler-Toscano, F., Rueda-Toicen, A., & Tegnér, J. (2018). A decomposition method for global evaluation of Shannon entropy and local estimations of algorithmic complexity. *Entropy*, *20*(8). https://doi.org/10.3390/e20080605

Ziv, J., & Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, *23*(3), 337–343. https://doi.org/10.1109/TIT.1977.1055714

Eidesstattliche Erklärung

Eidesstattliche Versicherung gemäß § 5 der Promotionsordnung vom 15.06.2018 der Mathematisch-Naturwissenschaftlichen Fakultät der Heinrich-Heine-Universität Düsseldorf:

Ich versichere an Eides Statt, dass die Dissertation „Methoden zur Messung und Modellierung des Zufalls in menschengenerierten Zahlensequenzen“ von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist. Ferner versichere ich, dass die Arbeit in der vorgelegten oder in ähnlicher Form bisher bei keiner anderen Fakultät als Dissertation eingereicht wurde und dass ich bisher keine erfolglosen Promotionsversuche unternommen habe.

Düsseldorf, den _____ _____

Datum Tim Angelike

Anhang: Einzelarbeiten

Originalartikel zu Studie 1:

Angelike, T., & Musch, J. (2023). *A comparative evaluation of measures to assess randomness in human-generated sequences* [manuscript submitted for publication]

Ich bin der Erstautor dieses Manuskripts. Ich war für die Planung, Programmierung und Auswertung der Studie verantwortlich. Ich habe außerdem das Schreiben des Manuskripts übernommen.

Originalartikel zu Studie 2:

Angelike, T., & Musch, J. (2023). *An improved modeling approach to investigate biases in human random number generation* [manuscript submitted for publication].

Ich bin der Erstautor dieses Manuskripts. Ich war für die Planung, Programmierung und Auswertung der Studie verantwortlich. Ich habe außerdem das Schreiben des Manuskripts übernommen.

A Comparative Evaluation of Measures to Assess Randomness in Human-Generated Sequences

Tim Angelike & Jochen Musch

Institute of Experimental Psychology, Department of Psychological Assessment and Differential Psychology, Heinrich-Heine University Düsseldorf

Author Note

Correspondence concerning this article should be addressed to Tim Angelike, Heinrich-Heine University, Universitätsstraße 1, 40225 Düsseldorf, Email: tim.angelike@hhu.de

Acknowledgments

We would like to thank Sandra Tyralla and Martin Papenberg for their helpful comments on an earlier draft of the manuscript.

Abstract

Whether and how well people can behave randomly is of interest in many areas of psychological research. The ability to generate randomness is often investigated using random number generation (RNG) tasks, in which participants are asked to generate a sequence of numbers that is as random as possible. However, there is no consensus on how best to quantify the randomness of responses in human-generated sequences. Traditionally, psychologists have used measures of randomness that directly assess specific features of human behavior in RNG tasks, such as the tendency to avoid repetition or to systematically generate numbers that have not been generated in the recent choice history, a behavior known as cycling. Other disciplines have proposed measures of randomness that are based on a more rigorous mathematical foundation and are less restricted to specific features of randomness, such as algorithmic complexity. More recently, variants of these measures have been proposed to assess systematic patterns in short sequences. We report the first large-scale integrative study to compare measures of specific aspects of randomness with entropy-derived measures based on information theory and measures based on algorithmic complexity. We compare the ability of the different measures to discriminate between human-generated sequences and truly random sequences based on atmospheric noise, and provide a systematic analysis of how the usefulness of randomness measures is affected by sequence length. We conclude with recommendations that can guide the selection of appropriate measures of randomness in psychological research.

Keywords: randomness, human random number generation, algorithmic complexity, entropy

A Comparative Evaluation of Measures to Assess Randomness in Human-Generated Sequences

Psychologists have long been interested in the human ability to generate random-like sequences (Baddeley, 1966; Falk & Konold, 1997; Wagenaar, 1972). The basic consensus is that humans generally do not behave randomly but instead exhibit systematic patterns that make their decisions predictable (Bocharov et al., 2020; Schulz et al., 2012; Shteingart & Loewenstein, 2016). Previous studies have used various measures from psychological research, computer science, and mathematics to quantify randomness or the lack thereof (Gauvrit et al., 2016; Ginsburg & Karpiuk, 1994; Oomens et al., 2015, 2021; Towse & Neil, 1998). However, the use of a plethora of measures has been criticized by researchers for being subjective and not allowing for comparisons between studies (Barbasz et al., 2008; Gauvrit et al., 2014, 2016), with the earliest criticisms dating back decades (Wagenaar, 1972). In this paper, we provide the first large-scale comparison of a diverse collection of randomness measures in terms of their ability to discriminate between random and human-generated sequences. We undertook this investigation to provide practitioners with data-based recommendations for selecting appropriate measures of randomness in psychological research.

The ability to generate random-like sequences is typically assessed using so-called random number generation (RNG) tasks, in which participants are asked to generate a sequence of random numbers. Typically, participants are asked to use numbers in the interval from 1 to 9 (Capone et al., 2014; Jokar & Mikaili, 2012; Miyake et al., 2000; Schulz et al., 2021; Zabelina et al., 2012), but there are also experimental paradigms that require the use of 10 or more numbers (Ginsburg & Karpiuk, 1994; Peters et al., 2007; Towse, 1998; Towse & Cheshire, 2007) as well as studies with only two numbers (Biesaga et al., 2021; Biesaga & Nowak, 2022; Gauvrit et al.,

2016). Another variant of this task requires participants to produce a random sequence of letters (Baddeley, 1966; Cooper et al., 2012; Larigauderie et al., 2020). Typically, participants are carefully instructed about the properties of random sequences to avoid measuring participants' misconceptions about randomness (Schulz et al., 2021; Towse & Cheshire, 2007).

The study of human RNG is of interest in several areas of psychology because it can be used to better understand various cognitive functions. For example, RNG performance can be used as an indicator of working memory and inhibitory capacity (Friedman & Miyake, 2004; Heuer et al., 2005) or, more generally, of the central executive (Cooper, 2016; Miyake et al., 2000; Miyake, Friedman, et al., 2001). Cognitive models have been developed to explain how humans attempt to generate random-like sequences of numbers (Cooper, 2016). Because RNG tasks are demanding and draw on different cognitive resources, they are often used as secondary load tasks to analyze performance on both primary and secondary tasks in dual-task experiments. Studies have shown a significant reduction in performance on the primary task and in the randomness of the sequences generated in the secondary load task (Howarth et al., 2016; Knott & Dewhurst, 2007; Miyake, Witzki, et al., 2001). RNG tasks have also been used to study the cognitive abilities of patients with psychiatric and neurological disorders such as schizophrenia (Peters et al., 2007; Shinba et al., 2000) or acquired brain injury (Maes et al., 2011), who show even more stereotyped behavior than healthy controls as evidenced by the tendency to generate series of adjacent number pairs (e.g., 7-6 or 4-5). The ability to generate random-like sequences has been found to develop similarly to other cognitive abilities across the lifespan: it increases from childhood to adolescence, peaking at age 25, followed by a decline that becomes steeper around age 60 (Gauvrit et al., 2017).

To study how different cognitive processes contribute to the ability to generate random-like sequences, researchers need measures that are sensitive to systematic patterns that people may exhibit. The problem that needs to be addressed is that, in principle, it is impossible to tell with certainty from a given sequence whether it was generated by a random or a deterministic process. An apparently systematic sequence such as 1-2-1-2-1-2-1-2 has the same probability of occurrence under a random process as the less systematic sequence 1-1-2-2-2-1-1-2 (Gauvrit et al., 2016). However, we can infer from the occurrence of systematic patterns in the first example sequence that it is more likely to have been generated by a deterministic, nonrandom process than the second sequence. Approaches to assessing the randomness of sequences generated in an RNG task are based on the identification of systematic patterns that provide evidence that a nonrandom process could have generated them (Gauvrit et al., 2016; Ginsburg & Karpiuk, 1994; Towse & Neil, 1998). Researchers have not agreed on a single way to assess randomness, as evidenced by the heterogeneous approaches summarized in the following section. Instead, the measures used to quantify randomness or the lack thereof, can be broadly grouped into three categories.

Approaches to the Detection of Deviations from Randomness

Measures Commonly Used in Psychological Research

Over the past several decades of research, psychologists have used a variety of measures to assess randomness. Many of these measures take into account biases in human behavior, such as the fear of repetition, which refers to the tendency to avoid direct repetitions of a number (Cooper, 2016). Two commonly used collections of randomness measures for analyzing human-generated random number sequences were proposed by Ginsburg and Karpiuk (1994) and Towse and Neil (1998). Ginsburg and Karpiuk proposed a collection of ten measures that assess typical

human biases in an RNG task: The coupon score, the gap score, the poker score, the runs index, the cluster ratio, the RNG index, diagram repetitions, repetitions, series, and variance of digits. These measures were often aggregated using principal component analyses based on the correlation between randomness measures across participants. These analyses revealed three components: cycling, seriation, and repetition. The cycling component is shown by the tendency to select numbers that have not been used recently with increased probability; the seriation component is shown by the tendency to stereotype behavior, such as the tendency to ascend or descend a series of numbers; and the repetition component reflects the avoidance of direct repetition (Peters et al., 2007).

Towse and Neil's work is a refinement of Ginsburg and Karpiuk's work, removing redundant measures, such as the poker score, the cluster ratio, diagram repetition, repetitions, and series, and adding new ones, such as the phi score - a measure of repetition that, unlike most of the measures discussed in Ginsburg and Karpiuk (1994), can be computed over any interval length -, the redundancy index, the turning point index, and the adjacency score. Towse and Neil also performed a principal component analysis on all of their measures. They concluded that the correlation between measures of randomness across participants was best explained by a four-component solution. Towse and Neil (1998) named these four components "equality of response usage", "short repetitions", "prepotent associates", and "long repetitions." Their "equality of response usage" component mirrors Ginsburg and Karpiuk's cycling component in that, both components reflect behavior that is aimed at using all numbers in a sequence in an equal manner. This is also evidenced by other randomness measures that load on these components, such as the coupon and the median gap score, which assess how long it takes for all numbers to occur once and how long it takes, on average, for a number to be repeated. The component "prepotent

associates" assesses whether there is a tendency to engage in stereotypical behavior, such as repeating some pairs of number more often than others in a sequence, which is conceptually similar to the seriation component. Randomness measures that assess the tendency to repeat number pairs load on both of these components in the Ginsburg and Karpiuk and Towse and Neil studies. The original repetition component of Ginsburg and Karpiuk (1994) was found in two components reflecting the number of repetitions over short and long distances. The generation of a number that has been used one to three numbers before in the sequence is an example of a short distance repetition; the generation of a number that has been used at least four numbers before is an example of a long distance repetition. The Towse and Neil collection comes with the RgCalc software tool and has recently been implemented in the R package randseqR (Oomens et al., 2021). The measures proposed by Towse and Neil are widely used in psychological research (Barbasz et al., 2008; Cooper, 2016; Larigauderie et al., 2020; Linschoten & Harvey Jr., 2004; Maes et al., 2011; Schulter et al., 2010; Zabelina et al., 2012). Some of the measures included in Towse and Neil's collection that assess whether some responses or pairs of responses are generated more frequently than others in a sequence, such as the redundancy index, the RNG index, and the RNG2 index, fall into a second category of entropy-based measures that can also be used to quantify the randomness of a sequence.

Block Entropy

As defined by Shannon (1948), entropy is a measure that quantifies the amount of information in a sequence based on the frequency distribution of the symbols that make up the sequence. A sequence in which each symbol is equally frequent has maximum entropy and contains the most information. A sequence consisting of only one symbol has an entropy of 0 and contains no information due to redundancy. When applied to the study of human RNG tasks,

entropy indicates whether there is an uneven distribution of the frequency of numbers in a sequence. In this context, a low entropy would indicate that a person has used some numbers more than others in a sequence. Block entropy is an extension of standard Shannon entropy that goes beyond the analysis of individual numbers by assessing whether there is an inequality in the frequency of blocks (also called *n*-grams) of consecutive responses in a sequence (Moore et al., 2018; Shannon, 1948). For example, a sequence might contain the block 7-4-1 more frequently than all other blocks of the same size, resulting in a lower block entropy. Some of the measures in Towse and Neil's (1998) collection that are used to assess whether there is an inequality of responses or pairs of responses in a sequence are variants of Shannon and block entropy. However, to our knowledge, measures of block entropy that assess whether there is an inequality of response triplets or quadruples (or even larger blocks) have not been systematically used to analyze human RNG tasks.

Measures of Algorithmic Complexity

Recently, some new promising measures of randomness have been proposed that originate from algorithmic complexity theory (Gauvrit et al., 2014; Zenil et al., 2018). These randomness measures are based on the Kolmogorov-Chaitin definition of complexity (Gauvrit et al., 2016). In this framework, complexity is defined as the length of the shortest computer program that can produce a given object, in this case, a sequence of numbers. A sequence that follows a systematic pattern, such as 1-2-1-2-1-2-1-2, can be generated by a program that follows a simple algorithmic rule (e.g., repeat response pair 1-2 four times). This sequence would therefore be considered not complex and probably not generated by a random process. However, the same sequence would be considered random according to classical Shannon entropy if, when choosing from two possible responses, both responses (e.g., 1 and 2) occur equally often. The

concept of algorithmic complexity is important for algorithms for lossless compression of long sequences, where the goal is to replace the original sequence with a shorter sequence that contains all of the information of the original sequence but replaces recurring patterns (Lempel & Ziv, 1976). A sequence is considered more random the less compressible it is (i.e., the fewer repeating patterns it contains). Measures of Lempel-Ziv complexity have only rarely been used in psychological research. However, the study by Wong et al. (2021) is an example of such an application.

The software needed to approximate the algorithmic complexity for short sequences has only recently been implemented (Gauvrit et al., 2014, 2016; Soler-Toscano et al., 2014). It is based on the coding theorem method, according to which sequences with low algorithmic complexity have a high probability of being produced by a deterministic process (in this case, a computer program), and sequences with high algorithmic complexity have a low probability of being produced by a deterministic process. The computer programs used to quantify the probability that a deterministic process produces a sequence are Turing machines. A Turing machine is a theoretical model of a general-purpose computer that produces an output based on a specified set of rules. By sampling from many Turing machines, it is possible to construct a frequency table showing how often a sequence is generated by a randomly drawn Turing machine and, thus, by a deterministic process. This information is then used to approximate the algorithmic complexity of a sequence by taking the negative logarithm of the relative frequency with which the sequence is generated. A higher algorithmic complexity indicates sequences that are rarely generated by a deterministic process, and a lower algorithmic complexity indicates less complex sequences that are frequently generated by a deterministic process. This approach promises to detect systematic patterns for short sequences ($\leq$12), which no previous measure of

randomness has been able to do. It has been extended to longer sequences by Zenil et al. (2018), who proposed the block decomposition method (BDM) by combining it with the concept of entropy, which allows penalizing the repeated use of identical blocks of numbers in a sequence. The usefulness of measures of algorithmic complexity has already been demonstrated for the analysis of binary sequences in a psychological setting (Biesaga et al., 2021; Biesaga & Nowak, 2022; Gauvrit et al., 2016), but not for sequences of more than two possible numbers.

The Present Study

We present the first large-scale comparison of a diverse collection of randomness measures from psychology (Ginsburg & Karpiuk, 1994; Oomens et al., 2021; Towse & Neil, 1998) and from information theory and algorithmic complexity theory (Gauvrit et al., 2016; Lempel & Ziv, 1976; Shannon, 1948; Zenil et al., 2018). We determine how well randomness measures are suited to determine whether a human or a random process generated a sequence. Thus, our approach focuses on identifying measures of randomness that best detect biases that humans exhibit when they attempt to generate random numbers. Which measures best detect evidence of systematic behavior that is not present in random sequences? The rationale behind this validation approach is that a measure that is sensitive to nonrandomness, i.e., systematic patterns exhibited by humans, should be able to discriminate between random and human-generated sequences with high confidence.

As a gold standard for comparison, we used data from atmospheric noise as a presumably truly random source (Furutsu & Ishida, 1961; Haahr, 2023). We chose atmospheric noise to avoid having to rely on computer-generated pseudorandom sequences based on deterministic algorithms as a validation criterion. Unlike pseudorandom number generators, true random

number generators such as atmospheric noise produce random numbers that are aperiodic and non-deterministic (Haahr, 2023).

In addition, we investigate how the length of a sequence affects the usefulness of measures to discriminate between human-generated and random sequences, as sequence lengths often vary widely across studies (Figurska et al., 2008; Ginsburg & Karpiuk, 1994; Schulz et al., 2021). Moreover, it is plausible to assume that the measures differ in their sensitivity to detect systematic patterns in human-generated sequences depending on the sequence length, as some measures, such as the complexity measure of Gauvrit et al. (2016), were specifically designed for short sequences, while other measures, such as compression algorithms, are usually only used to analyze longer sequences (Zenil et al., 2018). Based on our results, we try to derive practical recommendations for the selection of the most useful measures for the analysis of randomness in human behavior.

Methods

Design

Sequences of numbers were generated either by human participants in an RNG task or by continuously updated variations in the amplitude of atmospheric noise data accessed through an interface provided by the R package random (Eddelbuettel, 2017) from the website random.org (Haahr, 2023). For each measure of randomness, we computed the resampled correct classification rate at which sequences of numbers could be correctly assigned to their generating source. We used logistic regression models with each measure of randomness as the independent variable and the source of the sequences (human or random) as the dependent variable. Each logistic regression model was bootstrapped (n = 1000), meaning that the model was trained on 1660 randomly sampled sequences with replacement (equal to our sample size multiplied by 2, since there were as many random as human-generated sequences), and the model's predictive

performance was evaluated on the sequences that were not part of the training process. Going beyond previous approaches, we also present the first comprehensive investigation of how sequence length affects classification rates. To this end, we calculated the correct classification rate for complete sequences and the first 20, 50, and 100 digits of the 200-digit sequences.

Material

Random Number Generation Task

Instructions. Following the approach of Schulz et al. (2021) and Towse and Cheshire (2007), we instructed participants to consider the following essential features of random number generation: 1) equal probability of responses, 2) independence of responses from each other, 3) absence of patterns and unpredictability of responses. We explained the RNG task using the analogy of repeatedly drawing a number from 1 to 9 from a hat, returning the drawn number, and then shuffling the contents of the hat to repeat this procedure. We provided participants with examples of repetitive patterns to avoid and an example of what a random sequence might look like. We asked participants to generate a random number each time they heard a metronome tone. If they missed a response, they were instructed to move on and generate another number with the next sound.

Experimental Task. We used a 3 x 3 grid to record the responses, with a number from 1 to 9 displayed in each cell of the grid. The numbers 1, 2, and 3 were in the first row, 4, 5, and 6 were in the middle row, and 7, 8, and 9 were in the bottom row (in the order of their naming). The experimental paradigm of using a 3x3 grid for the RNG task was adopted from Maes et al. (2011), who showed that the use of this grid yielded similar results with respect to the widely used Towse & Neil (1998) collection of randomness measures as when the numbers were produced orally. Before the start of RNG task, participants had to confirm that they were ready to

do the task with the click of a button. Following this confirmation, the screen cleared for 2000 ms, after which a horizontally centered 3x3 grid of 450 x 450 px appeared. After a further 1000 ms, the rhythmic sound of a metronome began and was repeated every 1500 ms until the RNG task was completed. Participants were instructed to randomly select and click on one of the cells with their mouse each time they heard the metronome sound. In response, the selected cell changed its color to orange for 250 ms, providing visual feedback to the participant. For the 1000 ms following their selection, participants could not select another cell from the grid, ensuring that they could not speed through the experiment. Once the last trial of the task was completed, the grid disappeared, and the study moved to the next page. Participants had to complete 200 trials of the RNG task, which took exactly 5 minutes if they kept to the rhythm of the metronome. A green bar indicated their overall progress on the RNG task.

Procedure

The study was conducted using the online platform Unipark (https://www.unipark.com/). Participants were welcomed on the first page of the study. We informed them about the general purpose of the study (random number generation), their rights, and the intended use of their data in order to obtain their informed consent. On the following page, we asked for demographic information about their age, gender, German language proficiency, and educational level. Next, participants had to complete an audio check to ensure that they could hear the metronome during the RNG task. They had to listen to a short audio file in which they heard a rooster, and then choose which of several animals they had heard. On the next page, participants were given instructions on how to complete the RNG task. Instructions could only be skipped after 60 seconds, so clicking through the instructions was not possible. We then asked participants two simple multiple-choice questions to make sure they understood the instructions. The first

question asked how they should behave during the experiment (answer: randomly). The second question asked when participants should choose a random number (answer: at the sound of the metronome). Participants were excluded from the study at this stage if they answered one or both questions incorrectly. On the next two pages, participants were allowed to adjust the volume of the metronome so that the sound was comfortable, and then completed 10 test trials of the RNG task to get used to it. This phase was followed by the main experimental task, which consisted of 200 trials. After completing the experimental task, participants provided self-reports on a need for cognition scale (Lins de Holanda Coelho et al., 2020), a conscientiousness scale (Rammstedt et al., 2017), and on their mathematical abilities (e.g., school grade in mathematics, learning stochastics in school or at university). Additionally, participants could enter comments about this study. On the next page, they were asked to indicate whether they were serious about participating in the study (Aust et al., 2013), with assurances that this question would not result in forfeiting their compensation. On the last page, participants were debriefed and thanked for their participation. Participants could again enter comments in a text box for feedback, as they now knew the purpose of the study. The median time it took to complete the study was 12 minutes.

Sample

Participants were recruited through the online panel Bilendi (https://www.bilendi.de). Participants had to be at least 18 years old and be native German speakers or have a comparable language level to participate in the experiment. Participants could take part in the study using a desktop computer, a tablet (with a touchscreen) or a laptop. Another requirement of the study was that participants were able to play an audio stream on the device they were using to participate. Participants who did not correctly answer both comprehension questions about the

experimental paradigm on the first attempt (n = 181) were not allowed to continue at this stage due to presumed inattention, in order to ensure the quality of the data. The total sample consisted of 830 participants, as 21 participants had to be excluded for the following reasons: one participant reported using a 10-sided fair dice for the task; another participant used the same answer 50 times in a row (for a quarter of the task); 16 participants indicated in a seriousness check at the end of the study that they did not participate sincerely; one participant used only three of the nine possible numbers during the entire RNG task; two participants did not follow the rhythm of the metronome in the RNG task (median intertrial latency over 2000 ms). The final sample consisted of 405 men, 424 women, and one person who reported a non-binary gender. The age of the sample ranged from 18 to 87 years (M = 51.15, SD = 15.37). Most participants, 567, reported a certificate of secondary education or a high school diploma as their highest level of education, 245 participants had a college degree, 16 had obtained a Ph.D., and only two participants had not completed high school. The participants were compensated with 0.50 € for their participation in the study. To increase the motivation of the participants, we conducted a lottery and awarded an additional bonus of 5 € to the 30 participants who generated the most random sequences according to the coupon score (the participants did not know how we would determine the most random sequences). This was to provide an additional incentive for participants to be as random as possible in the RNG task. The lottery was announced at the end of the instructions for the RNG task.

Measures of Randomness

Commonly Used Measures in Psychological Research (Towse & Neil, 1998)

We computed the most widely used collection of randomness measures in psychological research, namely that of Towse and Neil (1998), who also described these measures in detail in

their review. To compute these measures, we used the R package randseqR as described in Oomens et al. (2021). For the computation, we used the *randseqR option* (same name as the package) as suggested by the authors of the package. Thus, for measures that rely on computing the frequency of occurrence of response pairs in a given sequence, the last response pair consisting of the last and (after starting over) the first sequence number was not considered.

Redundancy Index. The redundancy index is a measure of whether there is an inequality in the frequency of responses in a sequence, approaching 100 if a sequence consists of only one response and 0 for perfect equality of all possible responses. The redundancy index is a transformed version of the classical Shannon entropy.

Random Number Generation (RNG) Index. The RNG index measures whether pairs of responses in a sequence (e.g., 4-1, 1-5, and 5-6 in the sequence 4-1-5-6) are equally distributed, given the underlying frequency distribution of the first response in a pair. Thus, the RNG index is a measure of whether the transition probabilities from one response to another are equal. The index ranges from 0 (perfect equality of transition probabilities) to 100 (all transition probabilities are either 1 or 0).

RNG2 Index. The RNG2 index follows the same logic as the RNG index. However, instead of looking at the transition probabilities between consecutive responses, it computes whether there is an inequality in transition probabilities between interleaved responses by a gap of one. For example, in the sequence 4-1-5-6 the two pairs 4-5 and 1-6 are considered. The range of this index is identical to the regular RNG index.

Null-Score Quotient (NSQ). The NSQ is the proportion of response pairs that do not occur in a sequence relative to the number of possible response pairs. The measure is multiplied by 100 to obtain percentages. The range of this measure is from 0 to 100 with a value of 0

indicating that all possible response pairs occur in a sequence. For this measure, lower values indicate a more even distribution of response pairs and therefore a higher degree of randomness.

Coupon Score. The coupon score measures how long it takes for all possible responses in a sequence to occur. This measure is computed by iterating over a sequence and counting the time until each response has occurred at least once. The result is stored, and the procedure starts again with the first response after the completed set. The final score is the average of the lengths required to observe all of the responses. If a sequence does not contain all possible responses, the score is set to the length of the sequence + 1. Whether a particular score indicates low or high randomness can only be judged by comparing it to the average score of random sequences, because the average time it takes for all responses to appear depends strongly on the cardinality of the set of available numbers. For this comparison, we used the sequences based on atmospheric noise data, which we used as a benchmark for a random source of numbers.

Repetition Gap. The repetition gap is the average gap between identical responses in a sequence. We computed three variants of this measure: the mean, median, and mode over the distribution of gaps between identical responses. Like the coupon score, the repetition gap can only be interpreted as a measure of randomness when compared to the mean repetition gap of random sequences.

Adjacency Index. The adjacency index measures the proportion of ascending pairs relative to the total number of pairs in a sequence. The measure can be computed for ascending (e.g., 3-4) and descending pairs (e.g., 7-6), or a combination of both. The index is multiplied by 100 to represent percentages. On average, a random sequence will contain a percentage of adjacent pairs equal to the proportion computed by dividing the number of possible adjacent pairs by the number of all possible response pairs.

Turning Point Index. Turning points are defined as minima and maxima in a sequence (e.g., the sequence 1-3-5-4-3-7 has two turning points, 5 and 3). The number of observed turning points is then compared to the theoretically expected number of turning points and multiplied by 100. Values of random sequences for this measure range from 90 to 100 (Oomens et al., 2021). Higher values indicate more turning points than theoretically expected, and lower values indicate fewer turning points than theoretically expected.

Runs Index. The runs index computes the variance over the lengths of ascending subsequences in a sequence. For instance, the sequence 1-4-7-3-5 contains two runs, one of length 3 (1-4-7) and one of length 2 (3-5). The runs index is the variance computed over the two values 3 and 2, which represent the run lengths. The idea behind this measure is to capture the variability in the length of ascending subsequences. A higher value would indicate frequent switching between short and long runs of ascending numbers, and a value of 0 would indicate that all runs of ascending numbers in a sequence have the same length. This measure must also be compared to the expected value of randomly drawn sequences in order to interpret an observed value as random or not.

Phi Index. The phi index is a measure of repetitions of responses that are divided by a gap of other responses between them. More specifically, the measure counts the number of repetitions between the first and the last response of all blocks of specified length in a sequence and compares this frequency to the expected frequency of repetitions based on the observed number of repetitions between the first and last response of blocks that are one response smaller. Negative values indicate too few repetitons, and positive values indicate more repetitions than theoretically expected. We computed the phi index for blocks ranging in size from 2 to 10 to allow comparability with other measures computed over different block sizes.

Block Entropy

Block entropy is a measure that indicates whether there is an inequality of blocks of responses in a sequence. Blocks are determined by iterating over the sequence with a rolling window of size *k*. We defined *k* to be between 2 to 10, excluding only blocks of size 1 as the redundancy index in the previous section is a transformed version of Shannon's entropy (Shannon, 1948), which reduces to block entropy of size 1. High values of block entropy indicate an equal distribution of blocks of length *k*; low values indicate an inequality with a minimum of 0 indicating that a sequence consists of only one response. We chose block size 10 as a cutoff to allow comparison with complexity measures for short sequences, which only allow block sizes of up to 10 to be considered (see below).

Measures of Algorithmic Complexity

For clarity, we divided the group of algorithmic complexity measures into three subgroups: averaged algorithmic complexity measures for short sequences as proposed by Gauvrit et al. (2016), the block decomposition method as proposed by Zenil et al. (2018), and compression algorithms (Lempel & Ziv, 1976).

Averaged Algorithmic Complexity for Short Sequences. This measure was computed using the R-package *acss* by Gauvrit et al. (2016), which is based on the coding theorem method. Complexity was computed over a rolling window for each block of length *k* between 2 and 10. Block size 10 was chosen as the cutoff, because for block sizes 11 and 12, for nine possible values in a sequence, the complexity could not be computed for all possible sequences (Gauvrit et al., 2016). Finally, the mean of all complexity values was taken as an aggregate measure for the entire sequence.

Block Decomposition Method (BDM). The BDM (Zenil et al., 2018) was also computed over a rolling window for each block size *k* between 2 and 10. Each block was then assigned its algorithmic complexity from the previous section. However, instead of repeatedly counting the algorithmic complexity of recurring blocks, the total score is only increased by the logarithm of the frequency of a block after its first occurrence. A repetition of blocks is thus penalized by the BDM formula.

Compression Algorithms. We also computed two different compression algorithms: Lempel-Ziv complexity (Lempel & Ziv, 1976; LZ76) following the guidelines of Kaspar and Schuster (1987) and Dolan et al. (2018), and the gzip algorithm using the *memCompress*() function in the R programming language (R Core Team, 2023). The goal of compression algorithms is to search for repeating patterns in a sequence and replace them with a symbol representing that pattern. In this way, the length of a sequence can be reduced without losing information, since the original sequence can be reconstructed from the new compressed version of the sequence. This approach can also be used to test how random a sequence is, since random sequences without patterns should be difficult to compress, while systematic sequences with many repeating patterns should result in shorter compressed sequences.

Results

Data analysis was performed using the R environment for statistical computing version 4.3.0 (R Core Team, 2023). The following additional packages were used for the analysis: acss 0.3-2 (Gauvrit et al., 2016), randseqR 0.1.0 (Oomens et al., 2021), randfindR 0.1.0 (Angelike, 2022), papaja 0.1.1 (Aust & Barth, 2022), ggplot2 3.4.2 (Wickham, 2016), and ggpubr 0.6.0 (Kassambara, 2020). The data and code used in all analyses can be found at https://osf.io/xwzup/?view_only=1052e095327241d280e5602762a66f77.

Computation of Randomness Indices

First, we drew random sequences equal in length (200 digits) and number (830 participants) to the experimental data. For this purpose, we used the R package *random* (Eddelbuettel, 2017), which is an interface to the random.org website that generates data sequences of numbers based on atmospheric noise data (Haahr, 2023). Next, we computed all of the randomness measures summarized in the methods section over all sequences (human-generated and random). We also computed all measures over the first 20, 50, and 100 numbers of each sequence to examine the effect of sequence length on these measures.

Neither gender nor any of the personality variables showed a significant association with the measures of randomness under investigation, after a Bonferroni correction was applied to avoid inflating the alpha error. Participants who had studied stochastics at school or had received higher education tended to have higher randomness scores, even after a Bonferroni correction; but their level of randomness was still significantly lower than that of truly random data generated by atmospheric noise.

Classification Results

In this section, we investigated which measures of randomness were best suited to discriminate between human-generated and random sequences. For this purpose, we constructed logistic regression models for each individual measure of randomness where the score of a measure computed over all sequences was the independent variable, and the binary dependent variable was the source of generation of a sequence (human or random). The analysis was repeated for each randomness measure computed over all sequence lengths examined. To control for possible effects of overfitting, the correct classification rate of each model was determined through bootstrapping (n = 1000; further details are provided in the design section). This allowed

us to construct empirical confidence intervals of the correct classification rate for each randomness measure by selecting the 2.5th and the 97.5th percentiles of the bootstrapped correct classification rates. If the confidence intervals of two measures do not overlap, the difference between these two measures regarding the correct classification rate can be considered significant. Similarly, we considered a measure to show only approximate random performance if a confidence interval included the value .50.

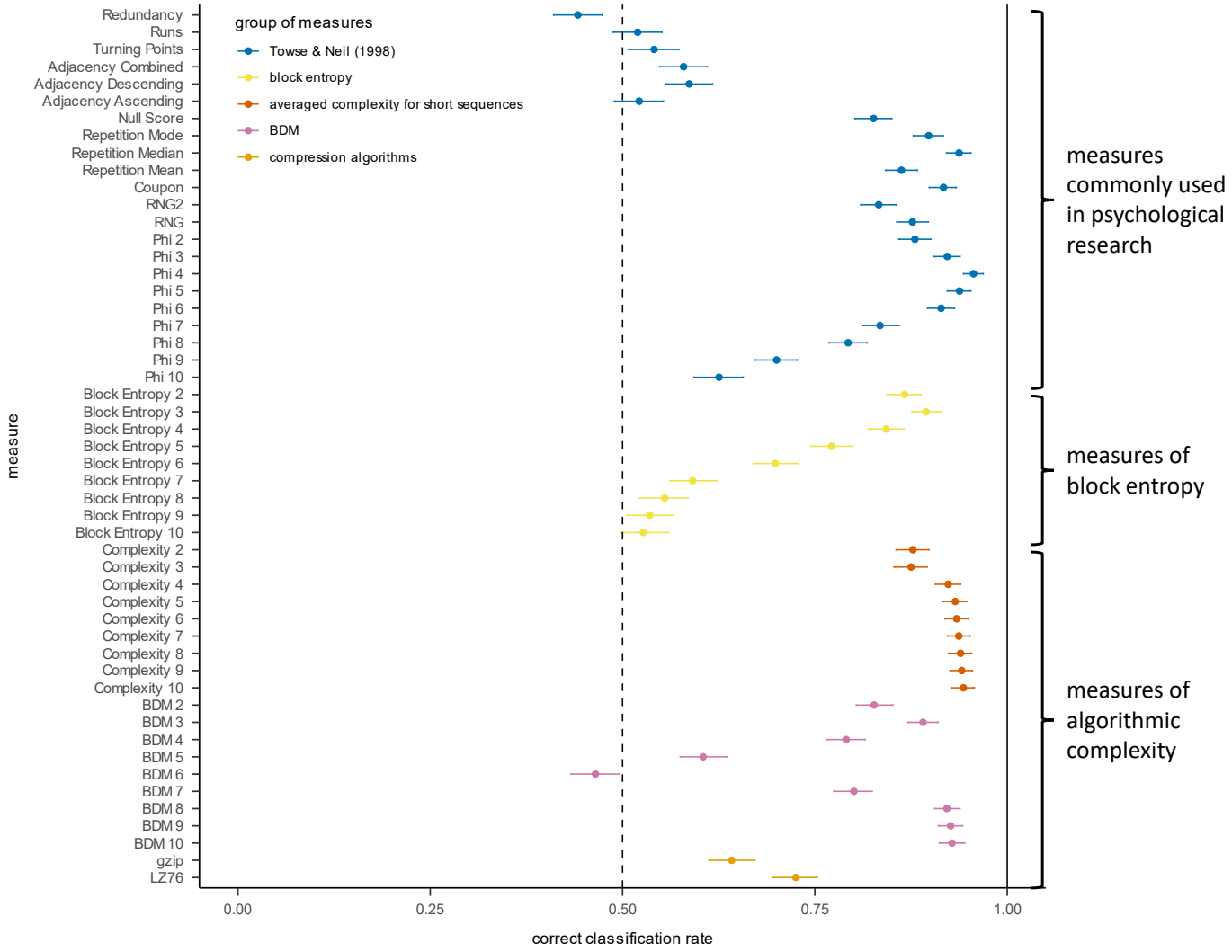
Overview of Measures

First, we provide an overview of the usefulness of the measures computed over the entire sequences in discriminating between human-generated and random sequences (see Figure 1 for visualization and the Appendix for all raw values). The correct classification rate was high for many randomness measures (M = 0.78 with a range from 0.44 to 0.96). Thus, on the basis of many randomness measures, it was possible to distinguish between human-generated and random sequences. Overall, measures of averaged algorithmic complexity were consistently useful for discriminating between the two sources of sequences (between .88 and .94), although the correct classification rate increased with block size. Similarly, the phi index showed a high correct classification rate (.63 to .96, highest for blocks of size 4). Interestingly, the results showed a higher range for the BDM and block entropy (.46 to .93 and .53 to .89, respectively). BDM measures showed high correct classification rates for larger block sizes of 7 to 10 (.80 to .93) and for smaller block sizes of 2 to 4 (.79 to 89) but not for moderate block sizes of 5 to 6 (.60 and .46). Block entropy measures were most useful for block sizes 2 to 4, with decreasing performance as block size length increased. Other useful measures for distinguishing between human-generated and random sequences were the coupon score (.92) and all variants of the repetition gap score, with the median gap between identical numbers being the most useful (.94).

The LZ76 showed significantly better than chance performance (.73) but fell short of other measures of algorithmic complexity. The runs-, redundancy-, and turning point index, as well as all variants of the adjacency index, showed, at best, slightly above-chance performance in distinguishing between human-generated and random sequences (.44 to .59).

Figure 1

Correct Classification Rate of Randomness Measures Ordered by their Group of Origin



Note. Correct classification rates for distinguishing between human-generated and random sequences using the randomness measures on the *y*-axis individually as predictors. The lines represent bootstrapped empirical confidence intervals (95%). The dashed line at $x = .50$ indicates

chance performance. The line at $x = 1.00$ indicates perfect performance. All measures can be assigned to one of three origins as indicated by the curly brackets on the right: measures commonly used in psychological research, measures of block entropy, and measures of algorithmic complexity (see methods section). The color coding indicates the type of the respective randomness measure. BDM = block decomposition method. LZ76 = Lempel-Ziv complexity. Numbers from 2 to 10 after the names of randomness measures indicate the block size that was used to compute the measure.

Groups of Randomness Measures

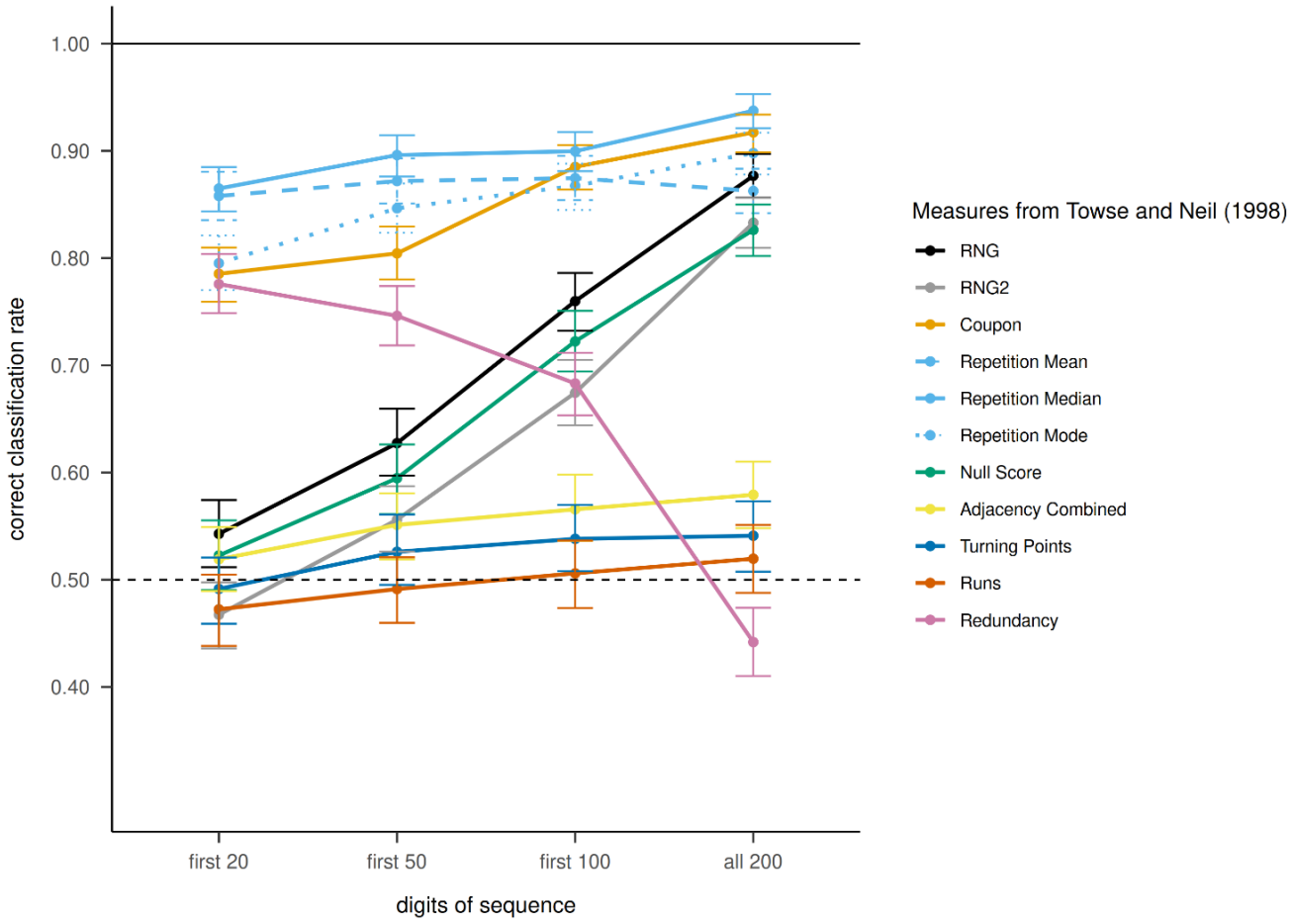
A complete collection of all randomness measures, including their descriptive values and correct classification rates by sequence length, can be found in the Appendix. This analysis is divided into five sections: measures commonly used in psychological research (Towse & Neil, 1998) using the R package randseqR (Oomens et al., 2021), block entropy measures, and complexity measures, with the latter being divided into measures of algorithmic complexity for short sequences, the BDM, and compression algorithms.

Common Measures in Psychological Research. The median repetition gap and the coupon score showed a high correct classification rate with only small effects of sequence length on performance. The mean and mode over the repetition gap between identical pairs were also among the most useful measures for distinguishing between human-generated and random sequences, although they performed slightly worse. The RNG and RNG2 indices and the NSQ showed only chance or near-chance performance for short sequences (size 20), but became increasingly useful for distinguishing between human-generated and random sequences as the sequence length increased. These measures assess systematic repetition in response pairs, and apparently require longer sequences to show clear differences between human-generated and

random sequences. The combined adjacency-, runs-, and turning point indices did not show high correct classification rates regardless of sequence length. One striking finding was that the redundancy index allowed adequate discrimination between human-generated and random sequences for the first 20 digits (.78), but this performance declined for longer sequences (.44 for the complete sequences). The redundancy index is a measure that assesses whether all possible responses (here, the numbers from 1 to 9) are equally likely to occur. In this experiment, human-generated sequences showed greater response equality than random sequences during the first 20 numbers of the sequence (see the Appendix). This effect disappeared in the long run, until there was no difference between the groups in the relative frequency of the numbers. Humans may show too much equality in the frequency of their responses, which is particularly evident for short sequences (Ginsburg, 1997). This observation is also consistent with previous findings that people may try too hard to use all responses equally compared to random sequences of the same length (Ginsburg & Karpiuk, 1994).

Figure 2

Correct Classification Rate into Human-Generated and Random Sequences of Logistic Regression Models Based on the Measures of Towse and Neil (1998)



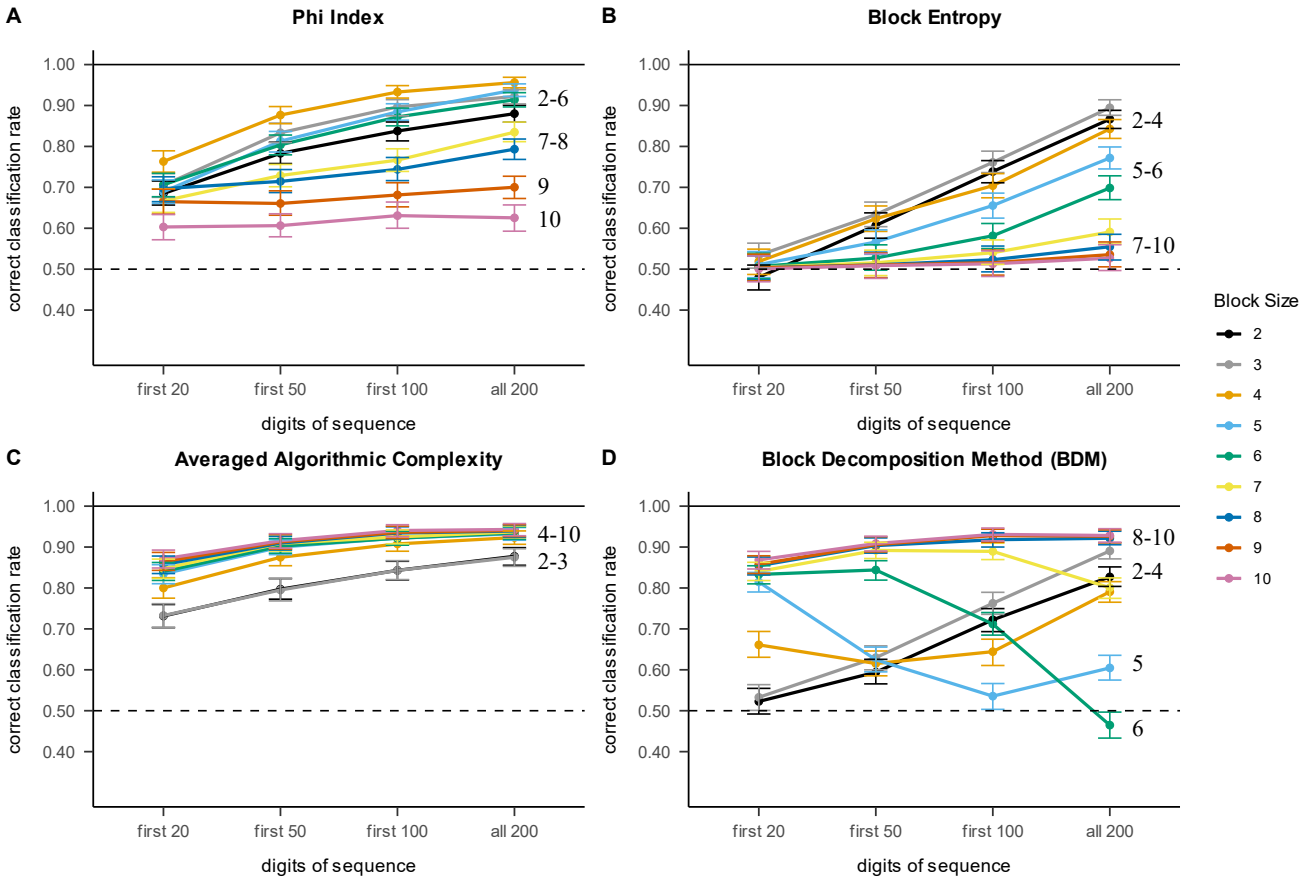
Note. Correct classification rates for distinguishing between human-generated and random sequences using the randomness measures in the legend as predictors. Error bars represent bootstrapped empirical confidence intervals (95%). The dashed line at $y = .50$ indicates the chance performance of a measure. The line at $y = 1.00$ indicates perfect performance. Due to the large number of randomness measures in this category (Towse & Neil, 1998), we excluded some less interesting measures from the visualization for the sake of clarity. In particular, we only included the combined adjacency score instead of using all of its variants because it included the information from the adjacency score for ascending and descending runs.

The correct classification rate obtained with the phi-index increased steadily the longer the sequence used for computing the measure. Performance was particularly high when computing the measure for blocks of size 4, although the correct classification rate was also high

for block sizes 2 to 6 (see Figure 3A). The correct classification rate was not as high for larger block sizes.

Figure 3

Correct Classification Rate into Human-Generated and Random Sequences of Logistic Regression Models Based on Measures Computed over Block Sizes 2 to 10



Note. Correct classification rates for distinguishing between human-generated and random sequences using different classes of randomness measures. The block sizes used to compute the measures are shown to the right of the corresponding lines and in the legend. Error bars represent bootstrapped empirical confidence intervals (95%). The dashed lines at $y = .50$ indicate the chance performance of a measure with respect to the correct classification rate. The lines at $y = 1.00$ indicates perfect performance.

Block Entropy. When examining block entropy, two interesting results can be highlighted (see Figure 3B). First, all measures of block entropy were uninformative for analyzing short sequences, as indicated by the near-chance performance in distinguishing between human-generated and random sequences using only the first 20 numbers of a sequence. Second, increasing the length of the sequence used to compute block entropy increased the correct classification rate into human-generated and random sequences. However, the magnitude of this increase seemed to depend on the block size. A clear increase in the correct classification rate can be seen for block sizes of 2 to 4. Block sizes of 5 to 6 yielded moderate increases in the correct classification rate, while block sizes of 7 to 10 were only marginally informative in terms of distinguishing between human-generated and random sequences for any given length of a sequence. This was probably due to the exponentially increasing number of distinct blocks with increasing size (9^k where k is the block size). Measures of block entropy look for an inequality in the use of blocks of a given size, which is particularly hard to find when the number of possible blocks is too large. Block entropy measures for large block sizes (e.g., 9 or 10) are likely to require much longer sequences to be informative.

Measures of Algorithmic Complexity. The section on measures of algorithmic complexity is divided into three sections: measures of algorithmic complexity for short sequences, the BDM, and compression algorithms.

Averaged Algorithmic Complexity for Short Sequences. In this section, we investigate the measures of algorithmic complexity as proposed by Gauvrit et al. (2016). Figure 3C shows that the correct classification rate of all measures of averaged algorithmic complexity increases steadily with the length of the sequence considered and the block size used to compute the measure. Overall, measures of averaged algorithmic complexity showed a consistently high

correct classification rate compared to all other measures with the lowest correct classification rate above 70% and the highest above 90%.

Note that we observed a rather surprising result for all measures of averaged algorithmic complexity. Higher values are associated with more complex and random sequences. However, we found that human-generated sequences had higher values for these measures than sequences generated by a random process (see Appendix). This finding seems to be caused by the short block length of numbers for which it can be computed. In this study, the measure was computed for blocks up to size 10. A high-complexity sequence usually contains all possible numbers equally often, as can be seen in the examples of high-complexity binary strings in Gauvrit et al. (2014). However, if there are nine possible numbers, high-complexity sequences will appear as if someone had cycled through all available numbers when generating the sequence, because each number will occur approximately once. As a result, short sequences of high complexity may resemble those generated by humans, who show such a cycling tendency in their behavior. This is illustrated by the negative correlation between algorithmic complexity for block size 10 and the coupon score, $r(828) = -.59$, $p < .001$, showing that participants with a stronger tendency to cycle through all available numbers (lower coupon score) obtain higher values of algorithmic complexity.

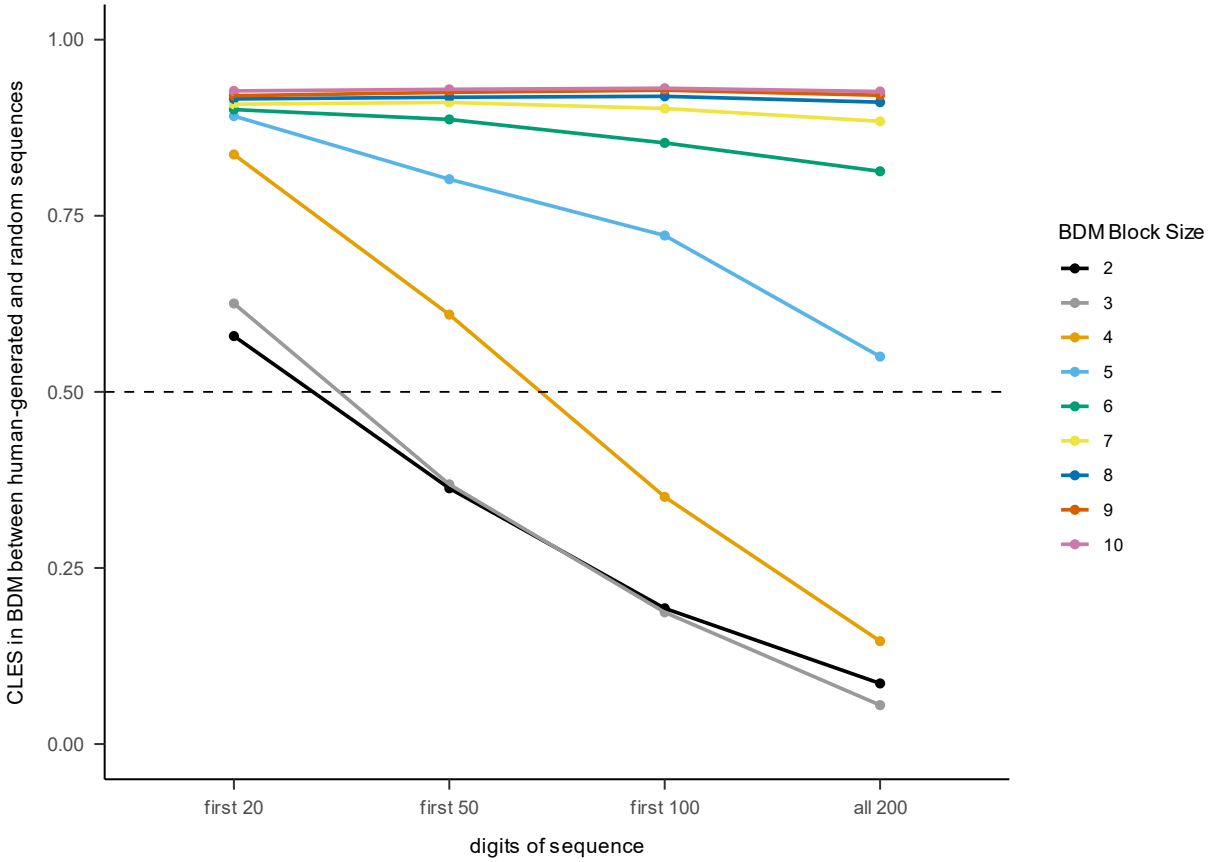
BDM. Results for BDM did not exactly follow the pattern of the averaged algorithmic complexity (see Figure 3D). For large block sizes (8 to 10), the BDM was useful for distinguishing between human-generated and random sequences. However, for block sizes 5 to 6, the correct classification rate decreased steadily with increasing sequence length. On the other hand, for block sizes 2 to 4, the correct classification rate using the BDM was at chance level

when computed over the first 20 numbers of the sequence but steadily increased with longer sequences.

To investigate this finding further, we computed the common language effect size for the difference in BDM scores between human-generated and random sequences, which indicates the probability that a randomly selected BDM score from the human sample is higher than a randomly selected BDM score from the random sequence sample (Figure 4). The results showed a general tendency: as sequence length increased, the differences between BDM scores for human-generated and random sequences decreased. For block sizes 2 to 4, this difference even reversed, so that random sequences had higher BDM scores than human-generated sequences. For larger block sizes, however, humans had consistently higher BDM scores regardless of sequence length. The reason for this pattern of results probably lies in the combination of algorithmic complexity and entropy in the BDM. Humans generally show a tendency to cycle through all available numbers too quickly, leading to higher averaged algorithmic complexity on the one hand and a tendency to repeat smaller blocks of size 2 to 4 on the other hand. Penalizing the latter leads to lower BDM scores for human-generated sequences for smaller block sizes. However, penalizing repetitive patterns becomes increasingly ineffective with increasing block size due to the exponentially increasing number of distinct blocks, as explained in the section on block entropy. As a result, human-generated sequences have higher BDM scores than random sequences for larger block sizes and lower scores for smaller block sizes. For moderate block sizes (5 to 7), the combination of these opposing effects may explain the declining performance of the measure as these effects appear to cancel each other out, resulting in smaller differences in BDM scores between human-generated and random sequences.

Figure 4

Difference In BDM Scores Between Human-Generated and Random Sequences Measured by the Common Language Effect Size (CLES)



Note. CLES = common language effect size. The dashed line at $y = 0.50$ indicates no difference between groups. Values above 0.50 indicate higher BDM scores for human-generated sequences compared to random sequences. Values below 0.50 indicate higher BDM values for random sequences compared to human-generated sequences.

Compression Algorithms. Both compression algorithms, LZ76 and gzip, showed a comparatively low correct classification rate into human-generated and random sequences (see Appendix). The correct classification rate peaked at about 60 – 70%. This is significantly lower than the highest correct classification rate of randomness measures from each of the measure groups analyzed so far. There were several examples of measures (e.g., the RNG index, phi index, measures of block entropy, or algorithmic complexity) that exceeded the 80% or even

90% correct classification rate. This finding is not very surprising given that compression algorithms are typically used to quickly compress longer sequences, such as files, and not to analyze human-generated sequences of a few hundred digits or less (Gauvrit et al., 2016; Zenil et al., 2018).

Discussion

The present study is the first large-scale integrative comparison of a broad collection of different measures of randomness. We analyzed not only measures that are traditionally used in psychological research (Towse & Neil, 1998), but also classical measures from information theory, such as block entropy (Moore et al., 2018; Shannon, 1948), as well as measures of algorithmic complexity (Gauvrit et al., 2016; Lempel & Ziv, 1976; Zenil et al., 2018). In addition, we analyzed how the effectiveness of measures for identifying human behavior may depend on sequence length. We also proposed a classification-based approach to evaluate randomness measures in terms of their usefulness in identifying human behavior in RNG tasks. For this analysis, we did not rely on pseudorandomly generated numbers from a computer for comparison; instead, we used sequences from a random source that are aperiodic and non-deterministic (Haahr, 2023).

Our results show that several measures of randomness can distinguish between human-generated and random sequences with a high correct classification rate of > .80. This is not too surprising, given the large body of research showing that humans generally fail to behave randomly (Bocharov et al., 2020; Figurska et al., 2008; Ginsburg & Karpiuk, 1994; Montare, 1999; Schulz et al., 2021). However, some randomness measures were particularly goot at distinguishing between human-generated and random sequences. Complexity measures such as averaged algorithmic complexity for larger block sizes (especially block size 10), block entropy

for shorter to moderately long block sizes (especially block size 3), and the phi index for moderately long block sizes (especially block size 4) were most useful. The median repetition gap score and the coupon score also showed large differences between human-generated and random sequences. We argue that researchers who wish to use measures of randomness that are sensitive to systematic patterns typical of humans should use these measures.

It should be noted, however, that the sensitivity of randomness measures to systematic patterns, which are often generated by humans, depends on the length of the sequence over which the measures are computed. Measures such as algorithmic complexity for blocks of size 10 or the repetition median already showed large differences between human-generated and random sequences for short sequences of length 20, with a correct classification rate between human-generated and random sequences close to 90%. If researchers want to analyze short sequences or subsequences of longer human-generated sequences, they should use these two measures. The phi score (block size 4) and the coupon score were also sensitive to differences between human-generated and random sequences for short sequences, but to a lesser extent than the averaged algorithmic complexity and the repetition median.

Averaged algorithmic complexity, the median repetition gap, the phi score, and, to a lesser extent, the coupon score also showed a high correct classification rate for longer sequences, demonstrating their applicability in various contexts of RNG tasks. On the other hand, measures such as block entropy showed almost no difference at all between human-generated and random sequences of short length. These measures required sequences of length 100 or more to achieve a high correct classification rate for moderate block sizes. Even for sequences of length 100, the correct classification rate between human-generated and random sequences by block entropy could not exceed the classification rate of the average algorithmic

complexity for sequences of length 20. A similar effect was observed for the RNG and RNG2 indices, which are modified measures of block entropy for number pairs. Several studies using these measures were based on the analysis of human-generated sequences of length 100 (Friedman & Miyake, 2004; Ginsburg & Karpiuk, 1994; Maes et al., 2011; Miyake et al., 2000; Miyake, Friedman, et al., 2001; Peters et al., 2007; Towse, 1998; Zabelina et al., 2012). Thus, it would seem advisable to increase the sequence length of the RNG task if researchers plan to use these measures, as longer sequences appear to be required to exploit their full potential. Our results suggest that block entropy derived measures for short to medium block sizes should only be used for sequences of at least, but preferably more than, 100 digits.

Several measures were not useful for identifying systematic patterns observed in humans: the turning point, and the runs index showed little or no difference between human-generated and random sequences regardless of sequence length. We advise caution in using these measures in future research, as they may introduce irrelevant variance for characterizing human behavior in RNG tasks. However, this finding should be replicated in future research to determine its stability. Otherwise, our findings are consistent with previous studies such as Ginsburg and Karpiuk (1994), who found similar differences in measures such as coupon and the median repetition gap between human-generated and random-like sequences. Our results are also consistent with their finding that humans show a more even use of all possible numbers in a sequence than would be expected on average for a random sequence of the same short length (100 numbers; Ginsburg & Karpiuk, 1994).

One measure that we recommend not to use for the analysis of sequences consisting of numbers in the range of 1 to 9 is the block decomposition method (BDM) due to its inconsistent interpretation. Depending on the length of the sequence and the block size used to compute it, a

larger value may indicative either a randomly generated sequence or a human-generated sequence. This is likely due to the opposing effects of complexity and block entropy on scores in the BDM: on the one hand, we found that the complexity was generally higher for human-generated than for random sequences; on the other hand, the BDM formula penalizes repetitions in a similar way to block entropy, leading to lower scores for humans, especially for blocks of short to medium length. We therefore argue that researchers should use average algorithmic complexity if they wish to use complexity measures, as it consistently shows higher values for human-generated sequences than for random sequences.

A strong argument can be made against the use of compression algorithms. The investigated measures, LZ76 and the gzip algorithm, even when computed over the complete sequences, performed worse regarding the correct classification rate than the averaged algorithmic complexity as proposed by Gauvrit et al. (2016), even when the latter was computed using only the first 20 numbers of each sequence. Therefore, we cannot recommend the use of compression algorithms as measures of randomness.

How can the randomness measures considered in this investigation be used for applied research questions? Fortunately, implementations are available for all of the measures presented in this paper. The measures commonly used in psychological research proposed by Towse and Neil (1998) can be computed using either the computer program from their original publication RgCalc or the more recent implementation in the R package randseqR by Oomens et al. (2021). The algorithmic complexity of short sequences as well as the BDM can be computed using the R package acss by Gauvrit et al. (2016), who also provide an introduction and tutorial on how to use it. The BDM can also be computed using the online algorithmic complexity calculator https://complexity-calculator.com/ (Soler-Toscano et al., 2014; Zenil et al., 2018).

Implementations of block entropy and LZ76 can be found in the R package randfindR at the following link https://github.com/TImA97/randfindR (Angelike, 2022). The code used to compute all randomness measures in this investigation can be found at https://osf.io/xwzup/?view_only=1052e095327241d280e5602762a66f77.

Finally, we must discuss the surprising result concerning the averaged algorithmic complexity as a measure of randomness. We found that human-generated sequences yielded higher estimates of averaged algorithmic complexity than random sequences. In this study, participants generated sequences containing the numbers 1 through 9. A highly complex sequence must contain all possible values approximately equally often, leaving little or no room for repetition if the sequence is only 10 in length. Consequently, a highly complex sequence with nine alternatives is a sequence that appears to show a certain cycling tendency. We argue that algorithmic complexity, as proposed by Gauvrit et al. (2014; 2016) for sequences with nine possible alternatives, does not accurately reflect randomness, since systematic non-random biases lead to higher values of complexity. Rather, the measure of algorithmic complexity appears to be inversely related to randomness. This limitation of the measure in terms of its interpretation needs to be addressed in future research. However, it should be emphasized that the averaged algorithmic complexity showed a high sensitivity to systematic patterns that humans exhibited when attempting to generate random sequences, regardless of the sequence length, underscoring the usefulness of this measure for characterizing human behavior.

A common criticism of the state of the scientific literature on the analysis of randomness in human-generated sequences is that too many different measures of randomness are used (e.g., Gauvrit et al., 2016; Wagenaar, 1972). This makes it difficult, if not impossible, to compare the results of different studies. The goal of this paper is to inform researchers about the properties of

the randomness measures they employ in their research. We analyzed a diverse collection of randomness measures in terms of their sensitivity to systematic patterns that humans show when trying to generate random sequences of numbers: measures that are motivated by psychological research (Towse & Neil, 1998), measures of block entropy (Shannon, 1948) and measures of algorithmic complexity (Gauvrit et al., 2016; Lempel & Ziv, 1976; Zenil et al., 2018). We went beyond previous research not only in the number and variety of randomness measures evaluated, but also in the systematic analysis of the influence of the sequence length on the measures' sensitivity to systematic human behavior in RNG tasks. We showed that some measures, such as the turning point and the runs index, show only a negligible difference between human-generated and random sequences. We argue that not all of the measures proposed by Towse & Neil (1998) may be necessary for analyzing sequences from an RNG task. We found that measures such as the phi index for moderate block sizes (a measure of repetition over a number gap), the coupon score (a measure of the cycling tendency), the repetition gap score, the block entropy of shorter to moderate block sizes for longer sequences, and especially the averaged algorithmic complexity regardless of sequence length show high sensitivity to the patterns exhibited by humans in an RNG task. We hope these results help researchers to make more informed decisions about the choice of randomness measures for the analysis of RNG tasks. For a reasonably well-specified research question, only one or a few sensitive randomness measures may be sufficient, rather than a large collection of randomness measures that may contain uninformative measures.

There are many different examples of analyzing the randomness of human-generated sequences from RNG tasks. Randomization performance can be analyzed to compare different experimental groups, e.g., different levels of production speed (Towse, 1998), or quasi-

experimental groups, e.g., healthy vs. schizophrenic patients (Peters et al., 2007) or healthy controls vs. patients with acquired brain injury (Maes et al., 2011). Performance on RNG tasks has been recognized as a possible indicator of executive function (e.g., Cooper, 2016). Deterioration in this performance can, thus, be used to infer the effect of an experimental manipulation or to uncover correlates of psychiatric and neurological disorders on cognitive functions. For such purposes, it seems prudent to use measures that have been shown to be most sensitive to systematic human behavior. We hope that this study will help researchers choose the most appropriate measure of randomness for their research question. However, researchers should not be completely discouraged from using other measures of randomness if they can better answer a theoretically meaningful question. For example, Peters et al. found that patients with schizophrenia tend to respond to pairs of adjacent numbers (such as 8-7 or 1-2). This could be explicitly investigated using the adjacency score, although it did not show a high sensitivity to systematic patterns found in humans in this study.

In the present study, we did not examine measures of recurrence quantification analysis (Oomens et al., 2015, 2023) because they measure similar information to Towse and Neil's (1998) traditional meaures when aggregated across complete sequences, which is the premise of this investigation. Future research should further investigate the effectiveness of randomness measures in assessing changes in randomness over time in an RNG task. This has already been done for binary sequences using measures of algorithmic complexity (Biesaga et al., 2021; Biesaga & Nowak, 2022; Gauvrit et al., 2016). A comparison between recurrence quantification analysis and algorithmic complexity for binary sequences and sequences containing more values may be promising.

In summary, we have compared a large collection of randomness measures for their usefulness in distinguishing between human-generated and random sequences, thereby establishing a validation criterion for judging the usefulness of a measure for identifying human behavior. These results are directly applicable to psychological research using RNG tasks. We find that some of the commonly used randomness measures are insensitive to the differences between human-generated and random sequences and are, therefore, not informative for characterizing human behavior. We also show that the sensitivity of many randomness measures can strongly depend on the sequence length used for analysis. On the other hand, some measures, such as the algorithmic complexity or the repetition gap score, showed high sensitivity to patterns indicative of human behavior for both short and long sequences. Taken together, these results can help guide practitioners in selecting the measures of randomness that are most appropriate for their research question.

Declarations

Funding

This research was conducted at the University of Duesseldorf. No external funding was received for conducting this study.

Conflicts of Interest/Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Consent to Participate

Informed consent was obtained from all individual participants included in the study.

Consent for Publication

All participants provided informed consent regarding publishing their data.

Availability of Data

The datasets generated during and/or analyzed during the current study are available in the OSF repository, https://osf.io/xwzup/?view_only=1052e095327241d280e5602762a66f77.

Code Availability

The code used to analyze the data is available in the same OSF repository as the data, https://osf.io/xwzup/?view_only=1052e095327241d280e5602762a66f77.

Open Practices Statement

The data and materials for all experiments and analyses are available at https://osf.io/xwzup/?view_only=1052e095327241d280e5602762a66f77.

References

Angelike, T. (2022). *randfindR: Analysis of randomness in human generated sequences* [Computer software]. https://github.com/TImA97/randfindR

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown* [Computer software]. https://github.com/crsh/papaja

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535. https://doi.org/10.3758/s13428-012-0265-2

Baddeley, A. D. (1966). The capacity for generating information by randomization. *Quarterly Journal of Experimental Psychology*, *18*(2), 119–129. https://doi.org/10.1080/14640746608400019

Barbasz, J., Stettner, Z., Wierzchoń, M., Piotrowski, K. T., & Barbasz, A. (2008). How to estimate the randomness in random sequence generation tasks? *Polish Psychological Bulletin*, *39*(1), 42–46. https://doi.org/10.2478/v10059-008-0006-7

Biesaga, M., & Nowak, A. (2022). *The role of the working memory storage component in the random-likes series generation*. PsyArXiv. https://doi.org/10.31234/osf.io/bvjkw

Biesaga, M., Talaga, S., & Nowak, A. (2021). The effect of context and individual differences in human-generated randomness. *Cognitive Science*, *45*(12), e13072. https://doi.org/10.1111/cogs.13072

Bocharov, A., Freedman, M., Kemp, E., Roetteler, M., & Svore, K. M. (2020). *Predicting human-generated bitstreams using classical and quantum models*. arXiv. https://doi.org/10.48550/ARXIV.2004.04671

Capone, F., Capone, G., Ranieri, F., Di Pino, G., Oricchio, G., & Di Lazzaro, V. (2014). The effect of practice on random number generation task: A transcranial direct current stimulation study. *Neurobiology of Learning and Memory*, *114*, 51–57. https://doi.org/10.1016/j.nlm.2014.04.013

Cooper, R. P. (2016). Executive functions and the generation of "random" sequential responses: A computational account. *Journal of Mathematical Psychology*, *73*, 153–168. https://doi.org/10.1016/j.jmp.2016.06.002

Cooper, R. P., Wutke, K., & Davelaar, E. J. (2012). Differential contributions of set-shifting and monitoring to dual-task interference. *Quarterly Journal of Experimental Psychology*, *65*(3), 587–612. https://doi.org/10.1080/17470218.2011.629053

Dolan, D., Jensen, H. J., Mediano, P. A. M., Molina-Solana, M., Rajpal, H., Rosas, F., & Sloboda, J. A. (2018). The improvisational state of mind: A multidisciplinary study of an improvisatory approach to classical music repertoire performance. *Frontiers in Psychology*, *9*, 1341. https://doi.org/10.3389/fpsyg.2018.01341

Eddelbuettel, D. (2017). *random: True random numbers using RANDOM.ORG* [Computer software]. https://CRAN.R-project.org/package=random

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, *104*, 301–318. https://doi.org/10.1037/0033-295X.104.2.301

Figurska, M., Stańczyk, M., & Kulesza, K. (2008). Humans cannot consciously generate random numbers sequences: Polemic study. *Medical Hypotheses*, *70*(1), 182–185. https://doi.org/10.1016/j.mehy.2007.06.038

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General*, *133*(1), 101–135. https://doi.org/10.1037/0096-3445.133.1.101

Furutsu, K., & Ishida, T. (1961). On the theory of amplitude distribution of impulsive random noise. *Journal of Applied Physics*, *32*(7), 1206–1221. https://doi.org/10.1063/1.1736206

Gauvrit, N., Singmann, H., Soler-Toscano, F., & Zenil, H. (2016). Algorithmic complexity for psychology: A user-friendly implementation of the coding theorem method. *Behavior Research Methods*, *48*(1), 314–329. https://doi.org/10.3758/s13428-015-0574-3

Gauvrit, N., Zenil, H., Delahaye, J.-P., & Soler-Toscano, F. (2014). Algorithmic complexity for short binary strings applied to psychology: A primer. *Behavior Research Methods*, *46*(3), 732–744. https://doi.org/10.3758/s13428-013-0416-0

Gauvrit, N., Zenil, H., Soler-Toscano, F., Delahaye, J.-P., & Brugger, P. (2017). Human behavioral complexity peaks at age 25. *PLOS Computational Biology*, *13*(4), e1005408. https://doi.org/10.1371/journal.pcbi.1005408

Ginsburg, N. (1997). Randomness: The Error of the Equal-Entry Matrix. *Perceptual and Motor Skills*, *85*(3_suppl), 1481–1482. https://doi.org/10.2466/pms.1997.85.3f.1481

Ginsburg, N., & Karpiuk, P. (1994). Random generation: Analysis of the responses. *Perceptual and Motor Skills*, *79*(3), 1059–1067. https://doi.org/10.2466/pms.1994.79.3.1059

Haahr, M. (2023). *RANDOM.ORG: True random number service* [Computer software]. https://www.random.org

Heuer, H., Kohlisch, O., & Klein, W. (2005). The effects of total sleep deprivation on the generation of random sequences of key-presses, numbers and nouns. *Quarterly Journal*

of Experimental Psychology Section A: Human Experimental Psychology, *58*(2), 275–307. https://doi.org/10.1080/02724980343000855

Howarth, S., Handley, S. J., & Walsh, C. (2016). The logic-bias effect: The role of effortful processing in the resolution of belief–logic conflict. *Memory and Cognition*, *44*(2), 330–349. https://doi.org/10.3758/s13421-015-0555-x

Jokar, E., & Mikaili, M. (2012). Assessment of human random number generation for biometric verification. *Journal of Medical Signals and Sensors*, *2*(2), 82–87. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3632045/

Kaspar, F., & Schuster, H. G. (1987). Easily calculable measure for the complexity of spatiotemporal patterns. *Physical Review A*, *36*(2), 842–848. https://doi.org/10.1103/PhysRevA.36.842

Kassambara, A. (2020). *ggpubr: "ggplot2" based publication ready plots* [Computer software]. https://CRAN.R-project.org/package=ggpubr

Knott, L. M., & Dewhurst, S. A. (2007). The effects of divided attention at study and test on false recognition: A comparison of DRM and categorized lists. *Memory and Cognition*, *35*(8), 1954–1965. https://doi.org/10.3758/BF03192928

Larigauderie, P., Guignouard, C., & Olive, T. (2020). Proofreading by students: Implications of executive and non-executive components of working memory in the detection of phonological, orthographical, and grammatical errors. *Reading and Writing*, *33*(4), 1015–1036. https://doi.org/10.1007/s11145-019-10011-6

Lempel, A., & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on Information Theory*, *22*(1), 75–81. https://doi.org/10.1109/TIT.1976.1055501

Lins de Holanda Coelho, G., H. P. Hanel, P., & J. Wolf, L. (2020). The Very Efficient Assessment of Need for Cognition: Developing a Six-Item Version. *Assessment*, *27*(8), 1870–1885. https://doi.org/10.1177/1073191118793208

Linschoten, M. R., & Harvey Jr., L. O. (2004). Detecting malingerers by means of response-sequence analysis. *Perception and Psychophysics*, *66*(7), 1190–1201. https://doi.org/10.3758/BF03196845

Maes, J. H. R., Eling, P. A. T. M., Reelick, M. F., & Kessels, R. P. C. (2011). Assessing executive functioning: On the validity, reliability, and sensitivity of a click/point random number generation task in healthy adults and patients with cognitive decline. *Journal of Clinical and Experimental Neuropsychology*, *33*(3), 366–378. https://doi.org/10.1080/13803395.2010.524149

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*(4), 621–640. https://doi.org/10.1037/0096-3445.130.4.621

Miyake, A., Witzki, A. H., & Emerson, M. J. (2001). Field dependence-independence from a working memory perspective: A dual-task investigation of the Hidden Figures Test. *Memory*, *9*(4–6), 445–457. https://doi.org/10.1080/09658210143000029

Montare, A. (1999). A Reversed Turing Test of Human Random Number Generation. *Perceptual and Motor Skills*, *88*(1), 138–146. https://doi.org/10.2466/pms.1999.88.1.138

Moore, D. G., Valentini, G., Walker, S. I., & Levin, M. (2018). Inform: Efficient information-theoretic analysis of collective behaviors. *Frontiers in Robotics and AI*, *5*. https://doi.org/10.3389/frobt.2018.00060

Oomens, W., Maes, J. H. R., Hasselman, F., & Egger, J. I. M. (2015). A time series approach to random number generation: Using recurrence quantification analysis to capture executive behavior. *Frontiers in Human Neuroscience*, *9*(JUNE). https://doi.org/10.3389/fnhum.2015.00319

Oomens, W., Maes, J. H. R., Hasselman, F., & Egger, J. I. M. (2021). RandseqR: An R package for describing performance on the random number generation task. *Frontiers in Psychology*, *12*, 629012. https://doi.org/10.3389/fpsyg.2021.629012

Oomens, W., Maes, J. H. R., Hasselman, F., & Egger, J. I. M. (2023). A time-series perspective on executive functioning: The benefits of a dynamic approach to random number generation. *International Journal of Methods in Psychiatric Research*, *32*(2), e1945. https://doi.org/10.1002/mpr.1945

Peters, M., Giesbrecht, T., Jelicic, M., & Merckelbach, H. (2007). The random number generation task: Psychometric properties and normative data of an executive function task in a mixed sample. *Journal of the International Neuropsychological Society*, *13*(4), 626–634. https://doi.org/10.1017/S1355617707070786

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rammstedt, B., Kemper, C. J., Klein, M. C., Beierlein, C., & Kovaleva, A. (2017). A Short Scale for Assessing the Big Five Dimensions of Personality: 10 Item Big Five Inventory (BFI-10). *methods*, *data*, 17 Pages. https://doi.org/10.12758/MDA.2013.013

Schulter, G., Mittenecker, E., & Papousek, I. (2010). A computer program for testing and analyzing random generation behavior in normal and clinical samples: The Mittenecker pointing test. *Behavior Research Methods*, *42*(1), 333–341. https://doi.org/10.3758/BRM.42.1.333

Schulz, M.-A., Baier, S., Timmermann, B., Bzdok, D., & Witt, K. (2021). A cognitive fingerprint in human random number generation. *Scientific Reports*, *11*(1), 20217. https://doi.org/10.1038/s41598-021-98315-y

Schulz, M.-A., Schmalbach, B., Brugger, P., & Witt, K. (2012). Analysing humanly generated random number sequences: A pattern-based approach. *PLoS ONE*, *7*(7), e41531. https://doi.org/10.1371/journal.pone.0041531

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shinba, T., Shinozaki, T., Kariya, N., & Ebata, K. (2000). Random number generation deficit in schizophrenia characterized by oral vs written response modes. *Perceptual and Motor Skills*, *91*(4), 1091–1105. https://doi.org/10.2466/pms.2000.91.3f.1091

Shteingart, H., & Loewenstein, Y. (2016). Heterogeneous suppression of sequential effects in random sequence generation, but not in operant learning. *PLOS ONE*, *11*(8), e0157643. https://doi.org/10.1371/journal.pone.0157643

Soler-Toscano, F., Zenil, H., Delahaye, J.-P., & Gauvrit, N. (2014). Calculating Kolmogorov complexity from the output frequency distributions of small Turing machines. *PLOS ONE*, *9*(5), e96223. https://doi.org/10.1371/journal.pone.0096223

Towse, J. N. (1998). On random generation and the central executive of working memory. *British Journal of Psychology*, *89*(1), 77–101. https://doi.org/10.1111/j.2044-8295.1998.tb02674.x

Towse, J. N., & Cheshire, A. (2007). Random number generation and working memory. *European Journal of Cognitive Psychology*, *19*(3), 374–394. https://doi.org/10.1080/09541440600764570

Towse, J. N., & Neil, D. (1998). Analyzing human random generation behavior: A review of methods used and a computer program for describing performance. *Behavior Research Methods, Instruments, & Computers*, *30*(4), 583–591. https://doi.org/10.3758/BF03209475

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin*, *77*(1), 65–72. https://doi.org/10.1037/h0032060

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wong, A., Merholz, G., & Maoz, U. (2021). Characterizing human random-sequence generation in competitive and non-competitive environments using Lempel–Ziv complexity. *Scientific Reports*, *11*(1), 20662. https://doi.org/10.1038/s41598-021-99967-6

Zabelina, D. L., Robinson, M. D., Council, J. R., & Bresin, K. (2012). Patterning and nonpatterning in creative cognition: Insights from performance in a random number

generation task. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(2), 137–145. https://doi.org/10.1037/a0025452

Zenil, H., Hernández-Orozco, S., Kiani, N. A., Soler-Toscano, F., Rueda-Toicen, A., & Tegnér, J. (2018). A decomposition method for global evaluation of Shannon entropy and local estimations of algorithmic complexity. *Entropy*, *20*(8). https://doi.org/10.3390/e20080605

Appendix

Descriptive Statistics and Correct Classification Rates Between Human-Generated and Random Sequences

Measure	M_{human}	SD_{human}	M_{random}	SD_{random}	Correct classification rate
First 20 numbers of the sequences					
Block Entropy 2	4.04	0.21	4.04	0.15	.48 [.45, .51]
Block Entropy 3	4.12	0.16	4.15	0.05	.53 [.50, .56]
Block Entropy 4	4.07	0.14	4.08	0.02	.52 [.49, .55]
Block Entropy 5	3.99	0.13	4.00	0.00	.51 [.48, .54]
Block Entropy 6	3.90	0.12	3.91	0.00	.51 [.48, .54]
Block Entropy 7	3.80	0.11	3.81	0.00	.50 [.47, .54]
Block Entropy 8	3.70	0.10	3.70	0.00	.50 [.47, .53]
Block Entropy 9	3.58	0.10	3.58	0.00	.50 [.47, .54]
Block Entropy 10	3.46	0.09	3.46	0.00	.50 [.47, .53]
RNG	15.11	11.73	13.11	8.72	.54 [.51, .57]
RNG2	12.24	10.43	11.72	8.01	.47 [.44, .50]
Coupon	13.78	4.07	19.41	2.55	.79 [.76, .81]
Repetition Mean	7.26	1.22	4.99	0.93	.86 [.84, .88]
Repetition Median	7.02	1.61	4.07	1.22	.86 [.84, .88]
Repetition Mode	5.95	2.75	2.48	1.80	.80 [.77, .82]
Null Score	79.92	1.93	79.93	1.68	.52 [.49, .56]
Adjacency Asc	9.72	8.77	9.40	6.39	.49 [.46, .52]
Adjacency Desc	10.66	7.15	9.77	6.47	.52 [.49, .55]
Adjacency Combi	20.38	10.59	19.17	8.64	.52 [.49, .55]
Turning Points	93.29	19.83	93.91	15.20	.49 [.46, .52]
Runs	0.73	1.35	0.72	0.56	.47 [.44, .50]
Redundancy	4.31	4.09	10.34	5.11	.78 [.75, .80]
Phi 2	-3.45	2.64	-1.38	2.89	.68 [.66, .72]
Phi 3	-4.73	1.45	-2.97	2.48	.70 [.68, .73]
Phi 4	-4.70	1.60	-1.88	2.53	.76 [.74, .79]
Phi 5	-4.93	1.89	-2.96	2.36	.69 [.66, .72]
Phi 6	-4.60	2.23	-2.28	2.11	.71 [.68, .73]
Phi 7	-4.84	2.32	-3.17	1.97	.67 [.64, .70]
Phi 8	-4.41	2.31	-2.46	1.64	.70 [.67, .73]
Phi 9	-4.36	2.08	-2.84	1.55	.67 [.63, .70]
Phi 10	-3.27	1.76	-2.48	1.35	.60 [.57, .63]
LZ76	13.37	1.09	12.56	0.91	.65 [.62, .68]
gzip	27.45	1.24	27.67	0.79	.53 [.50, .56]
Complexity 2	7.93	0.00	7.93	0.00	.73 [.70, .76]

Complexity 3	11.89	0.02	11.87	0.02	.73 [.70, .76]
Complexity 4	16.22	0.09	16.11	0.07	.80 [.78, .83]
Complexity 5	20.59	0.16	20.37	0.13	.83 [.81, .86]
Complexity 6	24.96	0.25	24.61	0.21	.84 [.82, .86]
Complexity 7	29.37	0.38	28.85	0.30	.85 [.82, .87]
Complexity 8	33.80	0.52	33.08	0.41	.86 [.84, .88]
Complexity 9	38.26	0.68	37.30	0.54	.86 [.84, .89]
Complexity 10	42.67	0.85	41.45	0.67	.87 [.85, .89]
BDM 2	135.87	11.81	135.79	10.24	.52 [.49, .55]
BDM 3	209.56	11.82	211.38	5.57	.53 [.50, .56]
BDM 4	273.75	12.50	273.55	2.79	.66 [.63, .69]
BDM 5	328.14	13.28	325.84	2.21	.81 [.79, .84]
BDM 6	373.28	13.93	369.14	3.08	.83 [.81, .85]
BDM 7	410.25	14.42	403.87	4.21	.84 [.82, .86]
BDM 8	438.76	14.78	430.02	5.34	.85 [.83, .88]
BDM 9	458.58	15.01	447.55	6.43	.86 [.84, .88]
BDM 10	469.05	15.29	455.98	7.37	.87 [.85, .89]

First 50 numbers of the sequences

Block Entropy 2	4.99	0.22	5.09	0.12	.61 [.58, .64]
Block Entropy 3	5.44	0.20	5.52	0.06	.63 [.60, .66]
Block Entropy 4	5.50	0.16	5.55	0.02	.62 [.59, .65]
Block Entropy 5	5.50	0.14	5.52	0.01	.57 [.54, .60]
Block Entropy 6	5.47	0.12	5.49	0.00	.53 [.50, .56]
Block Entropy 7	5.45	0.10	5.46	0.00	.52 [.48, .55]
Block Entropy 8	5.42	0.09	5.43	0.00	.51 [.48, .54]
Block Entropy 9	5.38	0.08	5.39	0.00	.51 [.48, .54]
Block Entropy 10	5.35	0.08	5.36	0.00	.51 [.48, .54]
RNG	23.98	7.99	19.94	4.40	.63 [.60, .66]
RNG2	21.21	7.18	19.07	4.38	.56 [.53, .59]
Coupon	15.44	5.61	25.34	8.96	.80 [.78, .83]
Repetition Mean	8.40	0.56	7.27	0.58	.87 [.85, .89]
Repetition Median	7.79	1.11	5.38	0.97	.90 [.88, .91]
Repetition Mode	6.40	2.66	2.35	1.66	.85 [.82, .87]
Null Score	57.25	4.58	55.09	3.16	.59 [.56, .63]
Adjacency Asc	10.39	7.32	9.67	4.19	.51 [.48, .54]
Adjacency Desc	11.27	5.45	9.79	4.17	.56 [.52, .59]
Adjacency Combi	21.66	9.21	19.47	5.37	.55 [.52, .58]
Turning Points	92.18	15.87	94.47	9.37	.53 [.50, .56]
Runs	0.82	1.11	0.73	0.34	.49 [.46, .52]
Redundancy	2.01	2.46	3.82	1.90	.75 [.72, .77]
Phi 2	-3.05	2.48	-0.65	1.79	.78 [.76, .81]
Phi 3	-4.11	1.40	-1.35	1.63	.83 [.81, .86]
Phi 4	-4.25	1.29	-0.82	1.69	.88 [.86, .90]
Phi 5	-4.04	1.57	-1.34	1.60	.81 [.79, .84]

Phi 6	-3.55	1.78	-0.72	1.59	.80 [.78, .83]
Phi 7	-3.28	1.87	-1.33	1.49	.73 [.70, .76]
Phi 8	-2.72	1.90	-0.82	1.41	.71 [.69, .74]
Phi 9	-2.47	1.77	-1.20	1.31	.66 [.63, .69]
Phi 10	-1.85	1.60	-1.02	1.29	.61 [.58, .64]
LZ76	25.37	1.84	25.29	1.17	.54 [.51, .58]
gzip	44.12	1.71	43.74	1.03	.58 [.55, .61]
Complexity 2	7.93	0.00	7.93	0.00	.80 [.77, .82]
Complexity 3	11.89	0.02	11.87	0.01	.80 [.77, .82]
Complexity 4	16.22	0.08	16.11	0.05	.87 [.85, .90]
Complexity 5	20.58	0.15	20.36	0.08	.90 [.88, .92]
Complexity 6	24.94	0.23	24.60	0.13	.90 [.88, .92]
Complexity 7	29.34	0.34	28.84	0.18	.91 [.89, .93]
Complexity 8	33.76	0.46	33.06	0.25	.91 [.89, .93]
Complexity 9	38.20	0.59	37.27	0.32	.91 [.90, .93]
Complexity 10	42.59	0.74	41.43	0.39	.92 [.90, .93]
BDM 2	282.90	28.32	296.25	19.38	.59 [.57, .63]
BDM 3	533.51	42.42	552.70	15.29	.63 [.60, .66]
BDM 4	745.71	47.31	754.63	7.08	.62 [.59, .65]
BDM 5	936.46	51.10	936.31	4.39	.62 [.59, .66]
BDM 6	1114.48	53.97	1107.03	5.70	.84 [.82, .87]
BDM 7	1284.09	56.35	1268.76	8.07	.89 [.87, .91]
BDM 8	1445.45	58.86	1421.64	10.63	.90 [.89, .92]
BDM 9	1598.52	61.05	1565.49	13.33	.91 [.89, .93]
BDM 10	1740.93	63.05	1698.50	16.02	.91 [.89, .93]

First 100 numbers of the sequences

Block Entropy 2	5.48	0.24	5.66	0.09	.74 [.71, .77]
Block Entropy 3	6.33	0.28	6.49	0.05	.76 [.74, .79]
Block Entropy 4	6.51	0.25	6.59	0.02	.70 [.67, .73]
Block Entropy 5	6.54	0.22	6.58	0.01	.66 [.62, .69]
Block Entropy 6	6.54	0.21	6.57	0.00	.58 [.55, .61]
Block Entropy 7	6.53	0.19	6.55	0.00	.54 [.51, .57]
Block Entropy 8	6.52	0.18	6.54	0.00	.52 [.49, .56]
Block Entropy 9	6.50	0.17	6.52	0.00	.52 [.49, .55]
Block Entropy 10	6.49	0.16	6.51	0.00	.51 [.48, .54]
RNG	32.21	6.60	27.15	2.42	.76 [.73, .79]
RNG2	29.77	5.70	26.60	2.37	.67 [.64, .71]
Coupon	15.84	5.68	25.46	5.71	.89 [.86, .91]
Repetition Mean	8.72	0.28	8.19	0.28	.87 [.85, .90]
Repetition Median	7.96	0.89	5.85	0.73	.90 [.88, .92]
Repetition Mode	6.48	2.58	2.05	1.33	.87 [.84, .89]
Null Score	35.95	7.25	29.75	3.77	.72 [.69, .75]
Adjacency Asc	10.64	6.69	9.61	2.93	.53 [.50, .56]
Adjacency Desc	11.40	4.68	9.87	2.91	.57 [.53, .60]

Adjacency Combi	22.05	8.77	19.48	3.99	.57 [.54, .60]
Turning Points	92.00	14.76	94.53	6.39	.54 [.51, .57]
Runs	0.95	1.80	0.73	0.21	.51 [.47, .54]
Redundancy	1.29	1.38	1.88	0.91	.68 [.65, .71]
Phi 2	-2.88	2.59	-0.31	1.25	.84 [.81, .86]
Phi 3	-3.78	1.51	-0.66	1.20	.90 [.87, .91]
Phi 4	-4.01	1.15	-0.42	1.23	.93 [.92, .95]
Phi 5	-3.77	1.35	-0.71	1.20	.88 [.86, .90]
Phi 6	-3.24	1.53	-0.37	1.18	.87 [.85, .89]
Phi 7	-2.66	1.58	-0.72	1.14	.77 [.74, .79]
Phi 8	-2.16	1.63	-0.34	1.09	.74 [.72, .77]
Phi 9	-1.75	1.56	-0.65	1.00	.68 [.65, .71]
Phi 10	-1.18	1.31	-0.49	1.02	.63 [.60, .66]
LZ76	42.52	3.42	43.80	1.45	.59 [.56, .62]
gzip	65.96	3.07	66.14	0.93	.43 [.40, .46]
Complexity 2	7.93	0.00	7.93	0.00	.84 [.82, .87]
Complexity 3	11.89	0.02	11.87	0.01	.84 [.82, .87]
Complexity 4	16.22	0.08	16.10	0.03	.91 [.89, .93]
Complexity 5	20.58	0.15	20.36	0.06	.92 [.90, .94]
Complexity 6	24.93	0.23	24.59	0.09	.92 [.90, .94]
Complexity 7	29.32	0.34	28.83	0.13	.92 [.91, .94]
Complexity 8	33.74	0.47	33.05	0.17	.93 [.92, .95]
Complexity 9	38.17	0.60	37.26	0.22	.94 [.92, .95]
Complexity 10	42.55	0.74	41.41	0.27	.94 [.93, .95]
BDM 2	425.82	45.56	464.63	23.32	.72 [.69, .75]
BDM 3	1019.44	105.61	1091.99	28.89	.76 [.74, .79]
BDM 4	1510.86	127.10	1550.98	14.81	.64 [.61, .67]
BDM 5	1938.47	143.43	1952.85	7.69	.54 [.50, .57]
BDM 6	2339.94	158.56	2336.34	8.53	.71 [.68, .74]
BDM 7	2731.34	173.20	2709.74	11.93	.89 [.87, .91]
BDM 8	3113.62	186.98	3073.67	15.86	.92 [.90, .93]
BDM 9	3487.51	199.40	3427.98	20.08	.93 [.91, .94]
BDM 10	3848.12	210.31	3768.56	24.47	.93 [.92, .95]

Complete sequences (200 numbers)

Block Entropy 2	5.79	0.27	6.01	0.05	.87 [.84, .89]
Block Entropy 3	7.12	0.38	7.37	0.05	.89 [.88, .91]
Block Entropy 4	7.46	0.36	7.59	0.02	.84 [.82, .87]
Block Entropy 5	7.54	0.33	7.61	0.01	.77 [.74, .80]
Block Entropy 6	7.55	0.30	7.61	0.00	.70 [.67, .73]
Block Entropy 7	7.56	0.28	7.60	0.00	.59 [.56, .62]
Block Entropy 8	7.56	0.27	7.59	0.00	.55 [.52, .59]
Block Entropy 9	7.55	0.26	7.58	0.00	.54 [.51, .57]
Block Entropy 10	7.55	0.24	7.58	0.00	.53 [.50, .56]
RNG	40.76	5.71	35.87	1.12	.88 [.86, .90]

RNG2	38.89	4.80	35.53	1.12	.83 [.81, .86]
Coupon	16.00	4.15	25.46	4.01	.92 [.90, .93]
Repetition Mean	8.86	0.22	8.62	0.13	.86 [.84, .88]
Repetition Median	8.05	0.79	6.11	0.53	.94 [.92, .95]
Repetition Mode	6.50	2.52	1.73	1.01	.90 [.88, .92]
Null Score	17.90	8.93	8.64	2.87	.83 [.80, .85]
Adjacency Asc	10.73	6.32	9.70	2.09	.52 [.49, .55]
Adjacency Desc	11.38	4.06	9.96	2.09	.59 [.56, .62]
Adjacency Combi	22.11	8.41	19.67	2.85	.58 [.55, .61]
Turning Points	91.85	14.10	94.62	4.48	.54 [.51, .57]
Runs	1.05	3.22	0.73	0.15	.52 [.49, .55]
Redundancy	0.95	1.13	0.93	0.46	.44 [.41, .47]
Phi 2	-2.73	2.73	-0.17	0.89	.88 [.86, .90]
Phi 3	-3.56	1.59	-0.32	0.86	.92 [.90, .94]
Phi 4	-3.86	1.11	-0.24	0.89	.96 [.94, .97]
Phi 5	-3.54	1.26	-0.34	0.87	.94 [.92, .95]
Phi 6	-3.06	1.33	-0.16	0.84	.91 [.90, .93]
Phi 7	-2.38	1.34	-0.41	0.82	.83 [.81, .86]
Phi 8	-1.87	1.43	-0.14	0.82	.79 [.77, .82]
Phi 9	-1.31	1.45	-0.33	0.81	.70 [.67, .73]
Phi 10	-0.77	1.11	-0.22	0.78	.63 [.59, .66]
LZ76	72.83	6.64	76.75	1.75	.73 [.70, .75]
gzip	108.96	6.93	111.12	1.08	.64 [.61, .67]
Complexity 2	7.93	0.00	7.93	0.00	.88 [.86, .90]
Complexity 3	11.89	0.02	11.87	0.01	.88 [.85, .90]
Complexity 4	16.21	0.08	16.10	0.02	.92 [.91, .94]
Complexity 5	20.57	0.15	20.36	0.04	.93 [.92, .95]
Complexity 6	24.92	0.24	24.59	0.06	.93 [.92, .95]
Complexity 7	29.31	0.36	28.82	0.09	.94 [.92, .95]
Complexity 8	33.72	0.50	33.05	0.12	.94 [.92, .95]
Complexity 9	38.14	0.65	37.25	0.15	.94 [.93, .95]
Complexity 10	42.51	0.80	41.40	0.19	.94 [.93, .96]
BDM 2	555.89	57.84	615.61	18.06	.83 [.80, .85]
BDM 3	1843.90	228.74	2068.56	51.68	.89 [.87, .91]
BDM 4	2978.33	309.34	3127.96	28.85	.79 [.77, .82]
BDM 5	3914.87	356.00	3984.43	13.37	.60 [.57, .64]
BDM 6	4773.71	396.97	4795.12	12.46	.46 [.43, .50]
BDM 7	5611.89	437.44	5591.98	17.06	.80 [.77, .82]
BDM 8	6437.00	476.95	6377.92	22.82	.92 [.91, .94]
BDM 9	7253.01	513.77	7152.85	29.04	.93 [.91, .94]
BDM 10	8050.52	547.28	7908.17	35.52	.93 [.91, .94]

Note. Asc = Ascending. Desc = Descending. Combi = Combined. Numbers at the end of a measure indicate the block size used in the calculation. Values in square brackets indicate empirical confidence limits (95%).

An Improved Modeling Approach to Investigate Biases in Human Random Number Generation

Tim Angelike & Jochen Musch

Institute of Experimental Psychology, Department of Psychological Assessment and Differential Psychology, Heinrich-Heine University Düsseldorf

Author Note

Correspondence concerning this article should be addressed to Tim Angelike, Heinrich-Heine University, Universitätsstraße 1, 40225 Düsseldorf, Email: tim.angelike@hhu.de

Abstract

Yousif et al. (2022) proposed a computational model to investigate the processes and biases involved in human random number generation (RNG). Their original two-parameter model includes a repetition parameter and a side-switching parameter representing influences of the immediately preceding number on the choice of the next number. We propose two changes to the model. First, we replace the side-switching parameter with a more general and less task-dependent distance parameter, which accounts for the tendency to select subsequent numbers that tend to be either closer to or further away from the previous number on the selected response pad. Second, we extend the computational model by adding a third parameter to account for the human tendency to select subsequent numbers with greater probability the longer the respective number has not been previously selected, following the pattern of the well-known gambler's fallacy. This new "cycling" parameter takes into account the most recent and all previous selections. The generalized distance parameter, and particularly the new cycling parameter, improved the fit of the model to human-generated sequences and the rate of successful predictions of the next choice from 14.09% to 26.48%, significantly exceeding the expected chance value of 1/9 = 11.1%. Model-driven simulations also showed that the extended three-parameter model could better account for systematic patterns that can be observed in human RNG tasks. The improved model could be useful in many contexts where human biases in RNG tasks are analyzed.

Keywords: human random number generation, fear of repetition, computer simulation, computational modeling.

An Improved Modeling Approach for the Description of Biases in Human Random Number Generation

Researchers typically investigate the ability to generate random-like sequences of numbers using random number generation (RNG) tasks in which participants are asked to generate a random sequence of numbers (Ginsburg & Karpiuk, 1994; Peters et al., 2007; Towse & Neil, 1998). The general consensus from these studies is that humans encounter significant challenges when attempting to generate truly random number sequences, and exhibit various dependencies between subsequent responses when generating a random-like series of numbers (Bocharov et al., 2020; Figurska et al., 2008; Shteingart & Loewenstein, 2016). For example, people tend to avoid immediate repetitions of numbers and often exhibit a tendency to repeat specific pairs of numbers.

The present study tries to gain a better understanding of the cognitive processes underlying human deficiencies in random number generation. It builds on the work of Yousif et al. (2022) who introduced a computational approach to account for potential biases, involving two parameters that model human behavior in RNG tasks: a repetition parameter expressing the tendency to generate too few or too many repetitions and a side-switching parameter indicating the propensity to switch between higher and lower numbers. In the present contribution, we propose two modifications to Yousif et al.'s original model. First, we replace the somewhat domain-specific side-switching parameter with a more generalized distance parameter that characterizes the decision to choose a number as a function of its distance from the previously selected number. Second, we expand the computational model with a third parameter that accounts for the human tendency to cycle too rapidly through all possible numbers in a sequence.

Our testing of the proposed extended model demonstrates an improved fit to human-generated sequences in an RNG task compared to the original two-parameter model. Moreover, model simulations demonstrate that the new three-parameter model, including the distance and the cycling parameter, can readily account for various systematic patterns observed in human behavior in RNG tasks.

The ability to generate sequences of numbers that appear random is a subject of frequent investigation through the RNG task (Ginsburg & Karpiuk, 1994; Towse & Neil, 1998). In this task, participants are asked to generate a sequence of numbers within a discrete interval, typically ranging from 1 to 9, aiming for maximum randomness (Capone et al., 2014; Maes et al., 2011; Schulz et al., 2021). A common observation in this task is that humans often struggle to produce random-like sequences and exhibit serial correlations between consecutive responses (Bocharov et al., 2020; Figurska et al., 2008; Schulz et al., 2021; Shteingart & Loewenstein, 2016). One of the most noticeable errors individuals make when attempting to generate random number sequences is a significant reduction in the frequency of immediate repetitions of a number, as compared to what would be expected by chance in a truly random sequence (Cooper, 2016; Ginsburg & Karpiuk, 1994). The avoidance of such repetitions is so pronounced that their occurrence is often exceedingly rare. Studies have demonstrated that this bias remains consistent in humans, irrespective of variations in RNG task parameters, including production speed, the set of possible numbers for sequence generation, task modality, and cognitive load (Towse, 1998; Towse & Cheshire, 2007). A potential explanation for this bias may be that people generally perceive repeated numbers as less random than alternating ones (Falk & Konold, 1997; Nickerson, 2002). Additionally, humans exhibit a systematic tendency to repeat specific numerical patterns (Ginsburg & Karpiuk, 1994; Schulz et al., 2021; Towse, 1998; Towse & Neil,

1998). One example of such a systematic pattern is the inclination to switch from the lower (1,2,3,4) to the higher half (6,7,8,9) of numbers (e.g., 3-8, or vice versa, 8-3; Yousif et al., 2022).

The challenge of generating random numbers may arise from the involvement of multiple cognitive processes required to complete this task. There is evidence that RNG tasks require different cognitive functions, including working memory, monitoring, and inhibitory ability (Cooper, 2016; Friedman & Miyake, 2004; Sexton & Cooper, 2014; Towse & Cheshire, 2007). Notably, performance in RNG tasks is compromised in individuals with neurological and psychiatric disorders (Gauvrit et al., 2016). Schizophrenic patients, for example, often exhibit a strong inclination to repeat specific patterns, such as adjacent pairs (Peters et al., 2007). The relevance of understanding impairments in random number generation to the comprehension of clinical disorders underscores the importance of further exploring the human capacity to generate random numbers and the processes involved.

Recently, Yousif et al. (2022) proposed a cognitive model aimed at elucidating biases exhibited by humans when generating sequences of random numbers. Initially designed for RNG tasks involving the generation of sequences of numbers ranging from 1 to 9, the model can be generalized to any RNG task, regardless of the allowed number range. Yousif et al.'s model introduces two parameters representing biases in RNG tasks: a repetition parameter and a side-switching parameter. The repetition parameter, denoted as ϵ, captures the systematic tendency to produce too few or too many repetitions in a sequence. Typically, humans exhibit a tendency to show too few repetitions in RNG tasks (Ginsburg & Karpiuk, 1994; Towse, 1998). The side-switching parameter, denoted as η, reflects the inclination to switch from lower to higher numbers and vice versa in consecutive responses. Yousif et al. (2022) demonstrated the presence of side-switching behavior in human-generated sequences. Combining these two biases, the

probability $\sigma(z)_i$ of choosing the *i*-th number in the vector of all *K* possible numbers is expressed using the following formula:

$$\sigma(z)_i = \frac{e^{\epsilon \cdot r_i + \eta \cdot s_i}}{\sum_{j=1}^{K} e^{\epsilon \cdot r_j + \eta \cdot s_j}}. \quad (1)$$

In this equation, ϵ represents the repetition parameter and η represents the side-switching parameter. The repetition parameter influences the probability of choosing a number based on whether it is a repetition of the previous number. The model accounts for this tendency by multiplying the repetition parameter by the value r_i in the vector *r* (length *K*), which is coded as 1 at position *i* if the previous number in the sequence was the same number and 0 otherwise. This ensures that the repetition parameter only directly affects the probability of the number that would be a repetition of the previous number. Positive values in the repetition parameter indicate a tendency to repeat a number. Negative values indicate a tendency to avoid repetitions (fear of repetition). A parameter value of 0 indicates the absence of any bias.

The side-switching parameter (η) is associated with the tendency to switch between lower and higher numbers. For this parameter, the possible numbers from 1 to 9 are divided into two sides: lower numbers in the range from 1 to 4 and higher numbers in the range from 6 to 9 (each side consisting of four numbers). The number 5 is defined as the separator between the two sides and is considered to be neither a lower nor a higher number. The value of the side-switching parameter is multiplied by the value s_i of the vector *s* (length *K*), which is coded as 1 if the previous response was from the other side (e.g., if a 3 precedes a 9) and -1 if the previous response was from the same side (e.g., if a 3 precedes a 4). The value of the vector entry representing the number 5 is always coded as 0 as this number does not belong to either side. The entire vector is coded as 0 if the previously generated number was a 5, as a side-switch is by definition not possible in this case. Positive values in the side-switching parameter indicate a

tendency to switch from lower to higher and from higher to lower numbers (e.g., from 2 to 8 or from 9 to 3). Negative values indicate the tendency to stay on the same side (e.g., a 4 after a 3 or a 7 after a 9). Again, a value of 0 indicates no bias at all.

The nominator on the right-hand side of equation 1 is the weight associated with the selection of a number, which is the sum of the effects of the repetition and the side-switching parameter. To convert this weight for a number into a probability, it is divided by the sum of the weights of all the possible numbers that can produced (the denominator). This ensures that all probabilities $\sigma(z)_i$ lie between 0 and 1 and sum to 1.

Using modeling approaches such as that by Yousif et al. (2022), it is possible to formalize the biases that humans show when trying to generate random sequences as model parameters. In this way, Yousif et al. (2022) were able to directly estimate the putative latent variables behind RNG performance: in this case, the tendency to make or avoid repetitions, quantified by the repetition parameter, and the tendency to either switch from lower to higher numbers and vice versa, or to prefer to stay on one side of the number range, quantified by the side-switching parameter. Their model-based approach also allowed them to determine the fit of their model, which is useful for comparing models and testing whether or not to add a parameter. Finally, models of human number generation can be used to simulate sequences of numbers that should resemble human-generated sequences which Yousif et al. successfully used to test model predictions.

Our first goal was to make the original by Yousif et al. (2022) model less dependent on the format of the RNG task. Specifically, we replaced the side-switching parameter with a distance parameter that describes the selection of a number in an RNG task as a function of the distance between the previously selected number and the currently selected number. This means

that it is possible to describe systematic behavior in an RNG task where numbers close to the previously selected number (e.g., pairs such as 7-6) are generated with either increased or decreased probability. This parameter is conceptually related to the idea of the side-switching parameter, where numbers from two different sides (lower or higher side) are roughly related to larger distances and numbers from the same side to smaller distances. It was included to be able to account for the findings of Towse (1998) who demonstrated a tendency of human participants to generate adjacent pairs of numbers, which would result in negative values of the distance parameter reflecting higher probabilities for numbers that have a small distance to the previously generated number.

The new distance parameter has the advantage of being less dependent on the specifics of the RNG task, as it can be tailored to any possible order of the numbers that can be used to generate the sequence. This allows task-specific features to be included in the modeling process. For example, in the second experiment by Yousif et al., participants generated random numbers by clicking on a horizontal line. For this one-dimensional response format, the distance could be computed as the numerical difference between the numbers (e.g., the distance in the response pairs 4-7 and 7-4 would be 3). This procedure for calculating the distances can also be used for oral production formats of RNG tasks, which are commonly used as well (Ginsburg & Karpiuk, 1994; Peters et al., 2007; Schulz et al., 2021; Towse, 1998). One possible prediction for an oral production format would be that the distances between numbers are best represented by their numerical distance, which is equivalent to representing the numbers on a horizontal line. However, numbers in an RNG task can also be presented in a 3x3 matrix (Kee et al., 2013; Maes et al., 2011), which could be accounted for by calculating the distances between numbers in the two-dimensional space of possible numbers. A simple dichotomization into lower and higher

numbers is necessarily somewhat arbitrary as there are several ways on how to dichotomize a two-dimensional matrix of numbers (e.g., lower vs. upper or left vs. right side). Replacing the side-switching parameter with the distance parameter in the model, we get

$$\sigma(z)_i = \frac{e^{\epsilon \cdot r_i + \delta \cdot d_i}}{\sum_{j=1}^{K} e^{\epsilon \cdot r_j + \delta \cdot d_j}}. \quad (2)$$

The distance parameter δ is multiplied by the distance d_i in the vector d of length K, which encodes the distance of a number from the previously generated number. Positive values of the distance parameter lead to higher probabilities for numbers that have a large distance to the previously generated number, while negative values of the distance parameter lead to higher probabilities for numbers that have a smaller distance to the previously generated number. A parameter value of 0 indicates the absence of any systematic behavior.

However, previous research has found another peculiarity that is not accounted for in the original model by Yousif et al. (2022). Humans have been observed to exhibit a tendency to cycle too quickly through all available numbers in an RNG task (Ginsburg & Karpiuk, 1994; Maes et al., 2011; Peters et al., 2007; Towse & Neil, 1998). Cycling describes the behavior whereby people tend to generate numbers with a higher probability the longer they have not been used in a sequence. For example, if the number 6 has not been generated recently, it is more likely to be generated next in the sequence than a number that has already been generated recently. This tendency to cycle will also lead to faster completion of the full set of numbers on average, as an attempt is made to use all possible numbers as evenly as possible. This bias is related to the concept of the gambler's fallacy (Tversky & Kahneman, 1971), where people mistakenly believe that a number that has not come up for a long time is likely to come up next, even though the probability of each number remains the same in each round. To account for this tendency, we propose to extend the modified version of the two-parameter model by adding a

third parameter reflecting a cycling bias. A major advantage of an additional cycling parameter over the repetition, side-switching, and distance parameters is that it uses the entire prior choice history, not just the immediately preceding choice, to describe the subsequent choice. Therefore, the cycling parameter describes the generation of numbers in an RNG task on a more global scale than the two parameters of the original model. In this way, it should be possible to capture systematic patterns in human-generated sequences, such as repetitions of numbers separated by a gap of several other numbers, thereby going beyond simple first-order dependencies.

In the following, we will refer to the new cycling parameter as β. Positive values of this parameter indicate a tendency to cycle through all available numbers too quickly. Negative values indicate that the probability of selecting a number decreases the longer it has not been selected. The value 0 indicates the absence of any bias. Adding the third parameter to the model in equation 2, we get

$$\sigma(z)_i = \frac{e^{\epsilon \cdot r_i + \delta \cdot d_i + \beta \cdot g_i}}{\sum_{j=1}^{K} e^{\epsilon \cdot r_j + \delta \cdot d_j + \beta \cdot g_j}}. \quad (3)$$

Here, *g* is the gap vector of length *K,* which codes at each position *i* how long a number has not been used. Each time a number is selected, its gap value is set to 1, and the gap value of every other number is incremented by one. At the beginning of a sequence, each gap value is initialized to 1. The model assumes that the effect of the cycling parameter increases linearly with the number of times a number is not selected. However, assuming a linear trend would consider a gap increase from 2 to 4 as equivalent to a gap increase from 12 to 14, even though the latter would be a much smaller increase on a relative scale. We therefore also consider a modification of the model that scales the gap to the last occurrence of a number on a logarithmic scale. This expresses the idea that once the distance to the last occurrence of a number is already large, the

effect of increasing the distance further becomes smaller and eventually reaches an asymptote. This way, we get

$$\sigma(z)_i = \frac{e^{\epsilon \cdot r_i + \delta \cdot d_i + \beta \cdot \log_2 g_i}}{\sum_{j=1}^{K} e^{\epsilon \cdot r_j + \delta \cdot d_j + \beta \cdot \log_2 g_j}}. \quad (4)$$

The advantage of adding a cycling parameter is that it reflects a well-documented human bias in RNG tasks (Ginsburg & Karpiuk, 1994; Peters et al., 2007) and allows more complex dependencies to be uncovered. The repetition, side-switching, and distance parameters only consider influences of the immediately preceding number on the following number, whereas the cycling parameter describes the choice of the following number as a result of the complete prior history of choices. To test which model best describes the human-generated sequences produced in RNG tasks, we used the data from an RNG task in a previous study (Angelike & Musch, 2023).

Hypotheses

1. We hypothesized that the new distance and cycling parameters would improve model fit and allow us to better account for observed human behavior in an RNG task. This expectation was based on the finding that humans have been observed to prefer to generate pairs of adjacent numbers (Ginsburg & Karpiuk, 1994; Towse, 1998) and to show a tendency to cycle through all available numbers too quickly when attempting to generate random sequences (Ginsburg & Karpiuk, 1994; Peters et al., 2007).
2. We expected that the model would also to be able to account for systematic behavior that humans show in RNG tasks. This hypothesis is based on the findings by Yousif et al. (2022), who showed that simulated sequences of numbers based on the participants' fitted parameter values exhibited systematic features that were not explicitly modelled. For example, they calculated the number of direction switches

between ascending and descending subsequences of numbers. This analysis showed that the model-generated sequences had similar characteristics to the human-generated sequences in terms of this metric. We extended this approach by utilizing a more extensive range of randomness measures to assess systematic features in sequences to determine the effectiveness of our enhanced model in capturing higher-order properties compared to the two-parameter model without a cycling parameter. We used various measures that were identified as being highly sensitive to systematic patterns observed in human-generated sequences in a recent study (Angelike & Musch, 2023).

Methods

The data for our analyses was obtained from Angelike and Musch (2023). In their study, participants completed an RNG task through an online platform, whereby 200 consecutive numbers were generated between the range of 1 to 9 every 1.5 seconds, accompanied by a metronome sound. Participants selected a number by clicking on it in a 3x3 grid, resembling a telephone or ATM keypad. The first row consisted of numbers 1 to 3, the second row of numbers 4 to 6, and the third row of numbers 7 to 9. Prior to the experiment, participants were instructed about the concept of randomness. The sample comprised 830 participants. Additional information regarding the sample, experimental paradigm, and instructions is provided in the original paper. The dataset can be accessed through the following link https://osf.io/xwzup/?view_only=1052e095327241d280e5602762a66f77.

Results

We conducted all statistical analyses using the R programming language R 4.3.0 (R Core Team, 2023). Additional software packages employed for analysis were rstan 2.21.8 (Stan

Development Team, 2022), randseqR 0.1.0 (Oomens et al., 2021), randfindR 0.1.0 (Angelike, 2022), ggplot2 3.4.2, (Wickham, 2016), ggdist 3.3.0 (Kay, 2023), gghalves 0.1.4 (Tiedemann, 2022), ggpubr 0.6.0 (Kassambara, 2020), patchwork 1.1.2 (Pedersen, 2022), papaja 0.1.1 (Aust & Barth, 2022), BayesFactor 0.9.12-4.4 (Morey & Rouder, 2022) and effsize 0.8.1 (Torchiano, 2020). Additionally, we used the STAN programming language to obtain model estimates based on Bayesian hierarchical modeling, employing Hamiltonian Markov chain Monte Carlo simulations using the No-U-Turn sampler (Carpenter et al., 2017). The code used for all analyses can be found at https://osf.io/sygdu/?view_only=8c602343aad54fe9a429903d3d553ceb.

All models presented in the results section were estimated on four separate chains (the default of rstan) with 10,000 iterations each to obtain a measure of model convergence. We discarded the first 5,000 iterations as warm-up to prevent potential convergence issues during initialization. Parameter estimates were initialized with random values in the range -2 to +2 (the default of rstan).

Candidate Models

We compared a total four models: the original two-parameter model by Yousif et al. (2022; equation 1), the two-parameter model with a distance instead of a side-switching parameter (equation 2), where the distance between numbers was computed as the Euclidean distance between the numbers in the 3x3 matrix of the RNG task used in the experiment by Angelike and Musch (2023), the three-parameter model adding a cycling parameter with a linear scaling of the gap to the last occurrence of a number (equation 3), and the three-parameter model with a cycling parameter based on a logarithmic scaling of the gap (equation 4).

All parameters within a model were estimated jointly in a hierarchical framework. Individual parameters for each participant regarding a bias were assumed to follow a normal

distribution (the prior). For example, the original two-parameter model by Yousif et al. (2022) consists of the repetition (ϵ) and side-switching parameter (η), where individual parameters were assumed to follow a normal distribution with mean μ and standard deviation σ (the hyperparameters):

$$\epsilon_n \sim \mathcal{N}(\mu_\epsilon, \sigma_\epsilon^2) \tag{5.1}$$

$$\eta_n \sim \mathcal{N}(\mu_\eta, \sigma_\eta^2). \tag{5.2}$$

Here, the subscript *n* ranges from 1 to *N* (the sample size), identifying each participant. The subscripts for the means μ and variances σ^2 indicate that each parameter has its own mean and variance. Thus, we estimated both the individual parameters for each bias and their mean and variance to obtain parameter estimates at the group level, as can be seen in equation 5. This approach was employed in estimating all four models outlined in the equations 1-4.

The hyperparameters representing the means of the parameters were drawn from a uniform distribution ranging between -10 and 10. This cutoff was chosen as a simulation of a model-generated sequence of length 100,000 with extreme parameter values of -10 or +10 in the repetition parameter led to extreme response behavior comparable to that of participants who did not seriously participate in the task: The proportion of repetitions for a repetition parameter of -10 was 0 and 0.99961 for a repetition parameter of 10. Similar results were obtained when using these extreme parameter values for the other parameters. An uninformative prior was chosen for the hyperparameters of the means, given the absence of any prior information on parameter distribution. For the same reason, we specified a uniform prior ranging from 0 to 10 for the standard deviations of the normal distributions (square root of the variances). For all hyperparameters of the four models, the convergence diagnostic $\hat{R}$ was 1.00, indicating convergence of model estimates across all chains.

Model Selection

To assess the model's adequacy, we calculated the average probability of predicting upcoming numbers in a sequence using the fitted parameter values from the posterior distribution. This was done separately for each participant and each model. If the model assigned the highest probability to the number that was actually selected, this number was considered a correct prediction (hit) and assigned a value of 1; otherwise, it was considered a miss and assigned a value of 0. If there were multiple numbers with the highest probability and one of them was selected, the model's uncertainty in prediction was equally distributed among the candidate numbers that were assigned the highest probability. For example, if the model allocated an equal highest probability (25% each) of all possible numbers for the next selection to the numbers 6 and 8 and one of these numbers was picked, then the predictive accuracy for that choice was calculated to be 0.5. This corresponds to one divided by the number of alternatives that were assigned the highest probability. In this example, the model successfully predicted that either a 6 or an 8 would be chosen next, but showed no preference between these two alternatives. Finally, we calculated the average accuracy of all predictions (either 0, 1 or $1/n$, in the event of a tie among n numbers for the highest probability) for each participant and for each model given the posterior distribution of the parameters. In this way, we assessed the models' ability to forecast the subsequent response using all previously selected numbers.

First, we compared Yousif et al.'s (2022) original two-parameter model with our modified version, which uses a distance instead of the side-switching parameter. The mean predictive accuracy for the modified two-parameter model (M = 15.88%, SD = 5.56%) exceeded that of the original two-parameter model (M = 14.09%, SD = 4.42%) for 70.72% of the participants. A paired sample t-test confirmed that there was a rise in predictive performance

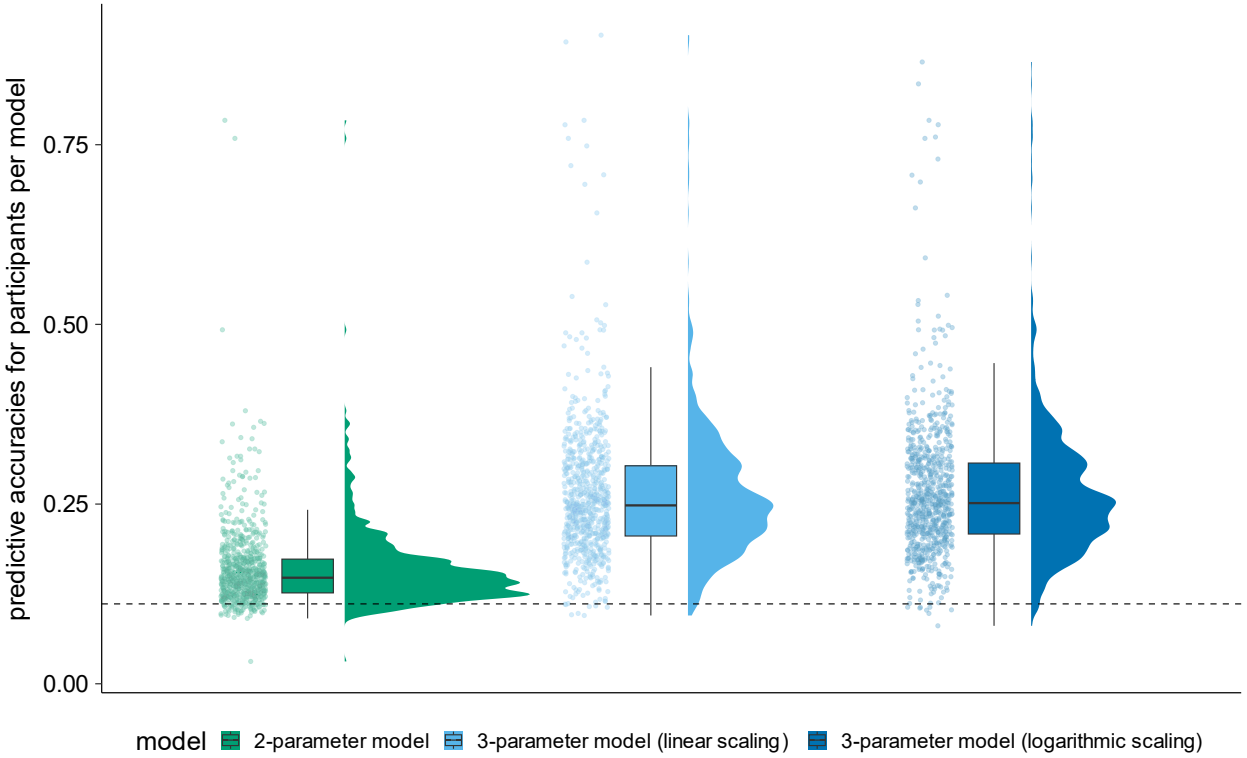
resulting from substituting the side-switching with the distance parameter, $t(829) = 16.78$, $p < .001$, $d = 0.34$, $\log_{10} BF = 51.07$. Cohen's d for dependent measures was calculated according to Borenstein et al. (2009).

It may seem surprising that both two-parameter models can only achieve a predictive accuracy marginally exceeding the chance level of 1/9 when anticipating the following number in a sequence. However, this outcome is explicable in light of the simplicity of the two-parameter models. The repetition, side switching, and distance parameters merely reflect simple first-order dependencies. Therefore, in cases where the repetition parameter has a strong negative influence, the model can ascertain that participants are inclined to eschew direct repetitions, but cannot determine the specific number that they will select next. Even in the best scenario, the inclusion of this parameter alone can therefore only increase predictive accuracy from 1/9 to 1/8, assuming that there are no repetitions present.

In the next step, we excluded the initial two-parameter model from further analysis as it was surpassed by the two-parameter model that replaced the side-switching parameter with a distance parameter. The three-parameter models were therefore built on the improved version of the two-parameter model as reflected in equation 3 and 4. The predictive probability for all participants regarding each of the three remaining models can be seen in Figure 1. This graph shows the increase in predictive accuracy achieved through the cycling parameter. To provide a comprehensive data presentation, raincloud plots were used for this and all other variable distribution visualizations (Allen et al., 2019).

Figure 1

Distribution of Predictive Accuracies for the Modified Two-Parameter Model Including a Distance Parameter, and the Two Three-Parameter Models Including a Cycling-Parameter Based on Either a Linear or Logarithmic Gap Distance.



Note. The distribution for each parameter is represented through three plots: a density plot (in arbitrary units), a boxplot, and a scatter plot. The bold line in the boxplot represents the median. The whiskers of the boxplot are limited to the IQR * 1.50. The jitter in the scatter plot along the *x*-axis was introduced to make all data points visible. The dashed line at $x = 1/9$ indicates chance performance in predicting the next number in a sequence.

Predictive accuracies of the three-parameter model employing a logarithmic scaling of the gap ($M = 26.48\%$, $SD = 9.31\%$) and the three-parameter model using linear scaling of the gap ($M = 26.14\%$, $SD = 9.29\%$) were better than for the two-parameter model with the distance parameter. When comparing the three-parameter model with logarithmic scaling of the gap and the two-parameter model, we found that the logarithmic cycling model predicted the subsequent

number in a sequence with higher accuracy for 95.06% of the participants. Moreover, there was a clear mean improvement in predictive accuracy when switching from the two-parameter model to the three-parameter model with a logarithmic scaling of the gap, $t(829) = 42.70$, $p < .001$, $d = 1.25$, $\log_{10} BF = 207.47$. Comparing the three-parameter model with the linear cycling of the gap with the modified two-parameter model, we found that for 94.22% of the participants the linear cycling model allowed for a better prediction of the next number in a sequence. This resulted in a significant increase in mean predictive accuracy, $t(829) = 41.00$, $p < .001$, $d = 1.22$, $\log_{10} BF = 197.53$. Thus, both the three-parameter model with the logarithmic scaling of the gap and the three-parameter model with the linear scaling of the gap outperformed the two-parameter model. These findings suggest that the addition of a cycling parameter to the model enhances its ability to predict choices in an RNG task. Taken together, the results are well in line with H1, which posits that adding a distance and cycling parameter would improve the model's fit.

When comparing the two three-parameter models with each other, it becomes apparent that the logarithmic cycling model achieved a higher predictive accuracy for 69.04% of the participants. Additionally, mean predictive accuracy was better under the logarithmic model compared to the linear three-parameter model, $t(829) = 10.24$, $p < .001$, $d = 0.04$, $\log_{10} BF = 19.96$. These findings suggest the presence of a cycling bias, wherein there is a higher likelihood of selecting a number if it has not been chosen recently. Even though the model with the logarithmic scaling of the gap showed a better fit, it should be noted that the absolute difference in predictive accuracy between the two three-parameter models was only relatively small.

Based on these results, we dropped the three-parameter model that assumed a linear scaling of the gap for the cycling parameter from further analyses and relied on the model with a

logarithmic scaling of the gap. However, we also retained the two-parameter model with the distance parameter in our subsequent analyses since we aimed to evaluate the effectiveness of the three-parameter model with a logarithmic scaling of the gap in capturing systematic patterns exhibited by humans. In the following analyses, we therefore compared the three-parameter model against the two-parameter model including the generalized distance parameter instead of the original side-switching parameter.

Simulation of Model-Driven Data

To assess how well our model can account for systematic patterns in human behavior in RNG tasks, we simulated model-generated sequences for the modified two-parameter model using the distance parameter and the three-parameter model using a logarithmic cycling of the gap. We conducted an assessment of both models' ability to account for biases observed in human behavior in RNG tasks. Based on the parameter values fitted for each model and each participant, we simulated 20 virtual participants with the same parameter values and the same sequence length (= 200). This resulted in 4,000 simulated numbers per participant and per model. We also created 20 permuted sequences for each participant to obtain random-like sequences composed of the same numbers as the original sequences (following the procedure proposed by Yousif et al., 2022). The permuted sequences were utilized as a benchmark for comparison, considering that the model-generated sequences ought to be more akin to human-generated sequences than random permutations of the same sequences.

Next, we calculated four measures of randomness across all sequences, including original human-generated, model-generated, and randomly permuted sequences. Our aim was to evaluate the efficacy of the two models in approximating human-generated sequences. In the same data set, the measures we employed were already shown to be sensitive to systematic behavior

exhibited by humans (Angelike & Musch, 2023). The first measure was the phi index for block size 4 (Oomens et al., 2021; Towse & Neil, 1998), a measure of interleaved repetitions. This measure indicates whether there are too many or too few repetitions when comparing the first and last number of blocks of size 4. For this metric, negative values denote insufficient repetitions, whereas positive values indicate an excess of repetitions. The second measure was block entropy for blocks of size 3 (Moore et al., 2018; Shannon, 1948), which is a measure that indicates whether there is an inequality in the frequency of blocks of responses with length 3 in a sequence. The value 0 indicates that only one block of numbers was used throughout; higher values indicate a more even distribution of blocks. The third measure was the coupon score (Ginsburg & Karpiuk, 1994), which calculates the average time taken for all numbers to occur at least once in a sequence. Lastly, we calculated the mean gap score (Towse & Neil, 1998), indicating the average distance between identical numbers in a sequence.

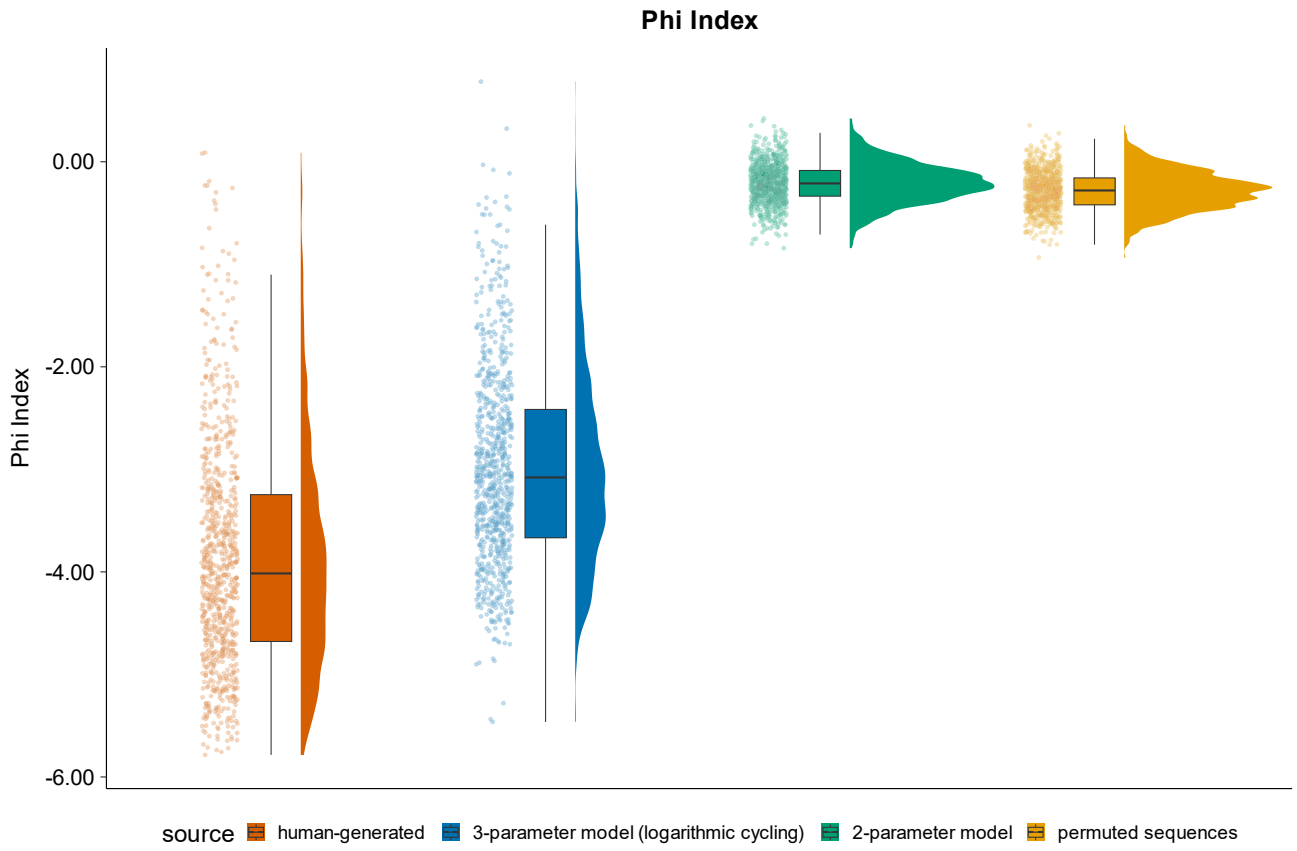
After the computation of the four randomness measures across all human-generated, model-generated, and permuted sequences, we computed for each participant the mean over the randomness measures over the respective participant's model-generated sequences as well as over the same participant's permuted sequences. This way, we obtained a mean model prediction for each model per participant for all four randomness measures as well as the same prediction for the randomly permuted sequences belonging to the same participant. If the models could account for the variance in the randomness measures under investigation, we would expect the mean model predictions for a randomness measure to be closer to the observed randomness measures of the participants than the randomness measures computed over the randomly permuted sequences. The distributions of the phi index from all data sources can be found in Figure 2. From a visual inspection, we see that the distribution of the phi index computed over

human-generated sequences (M = -3.86, SD = 1.11) can be best approximated by the sequences generated by the three-parameter model with the logarithmic scaling of the gap for the cycling parameter (M = -2.99, SD = 0.97). In comparison, the distribution of the phi index computed over sequences generated by the two-parameter model (M = -0.21, SD = 0.20) is no closer to the distribution of the human-generated sequences than the distribution of the randomly permuted sequences (M = -.28, SD = 0.19). This indicates that the two-parameter model without the cycling parameter cannot account for the lack of repetitions humans show over moderately long gaps (here between the first and last number of blocks of size 4), as can be seen by the negative phi index values in the human-generated sequences. This is not surprising given that the two-parameter model describes the decision for a number in a sequence only as a function of the number selected immediately before. In contrast, the cycling parameter describes this decision as a result of the complete prior choice history which explains why the three-parameter model can better account for systematic patterns between numbers that are not immediate neighbors.

For the sake of brevity, a full description of the distribution of the remaining three randomness measures for data from different sources is provided in the Appendix. The general pattern was similar for all randomness measures: the approximation of human-generated sequences concerning different measures of randomness was better for the three-parameter model with the logarithmic scaling of the gap for the cycling parameter than for the two-parameter model without a cycling parameter.

Figure 2

Distribution of the Phi Index over Sequences Generated by Human Participants, the Three-Parameter Model with a Logarithmic Scaling of the Gap, the Two-Parameter Model with a Distance Parameter and Random Permutations of Human-Generated Sequences



Note. This plot shows the distribution of the phi index computed over different sources of sequences: the human-generated sequences from a RNG task, sequences generated through the three-parameter model with the logarithmic scaling of the gap to the last occurrence of a number, the two-parameter model with a distance instead of a side-switching parameter and sequences that were random permutations of the original-human-generated sequences. The distribution for each source is represented in three plots: a density plot (in arbitrary units), a boxplot, and a scatter plot where the variation along the *x*-axis was introduced to allow better representation of the data points. The bold line in the boxplot represents the median. The whiskers of the boxplot are limited to the IQR * 1.50.

To assess the extent of disparity between the sequences that were formed by humans, by the competing models, or by randomly permuting human sequences, we adopted the classification methodology utilized by Angelike and Musch (2023). In doing so, we performed

non-parametric bootstrapping (n = 1000) to compute logistic regression models employing the randomness measure as the independent variable, and the data source as the dependent variable. This process was repeated for every combination of the human-generated sequences with the other sources (using a three-parameter model with logarithmic cycling, a two-parameter model with the distance parameter, and permuted sequences). For every bootstrapping iteration, we trained the logistic regression model by drawing 1660 samples with replacement from the computed randomness measures. This sample size was chosen to equal twice the human sample size as we had to pair each human-generated sequence with a sequence from one of the other sources. The correct classification rate for every bootstrapping iteration was established through prediction of the source of generation of the out-of-bag data, which was not part of model training. Thus, we acquired 1000 estimates of the correct classification rate, enabling us to calculate the average and the 2.5th and 97.5th empirical confidence intervals (as presented in square brackets). This methodology was reiterated for each measure of randomness. Correct classification rates of 50% indicate that a randomness measure could discriminate between human-generated and model-generated sequences only at chance level and thus, that the model-generated sequences very closely mimicked human-generated sequences.

The analysis of the phi index reveals that the disparity between human-generated and three-parameter model-generated sequences was smaller than that between human-generated and two-parameter model-generated sequences, as well as that between human-generated and randomly permuted sequences. Specifically, comparing human-generated sequences with sequences generated with the three-parameter-model using the phi index resulted in a correct classification rate of 67.78% [64.98%, 70.92%] in correctly identifying the generating source. Distinguishing between human-generated sequences and sequences generated with the two-

parameter-model was possible with a correct classification rate of 99.17% [98.64%, 99.68%], and distinguishing between human-generated sequences and sequences generated by permuting human-generated sequences was possible with a correct classification rate of 99.10% [98.50%, 99.67%]. The difference in the phi index between the human-generated sequences and the sequences generated by the two-parameter model was significant, as was the difference in the phi index between the human-generated and the permuted human-generated sequences. Taken together, the three-parameter model provided an improvement over the two-parameter model without a cycling parameter or no model at all (randomly permuted sequences). Nevertheless, a distinct dissimilarity was still noticed between the sequences produced by the three-parameter model and the ones created by humans, indicating that the three-parameter model did not approximate human-generated sequences perfectly. The findings obtained with three additional randomness measures (see Appendix) bolster this assertion: the three-parameter model was more successful in approximating organized structures manifested by humans in RNG tasks than the two-parameter model lacking a cycling parameter. The block entropy measure provided the only exception to this rule because regarding this measure, the three-parameter model was only marginally superior to the two-parameter model. These outcomes are well in accordance with H2 as model-driven sequences generated with the three-parameter model exhibited closer similarity to human-generated sequences than to randomly re-ordered human-generated sequences.

Parameter Estimates

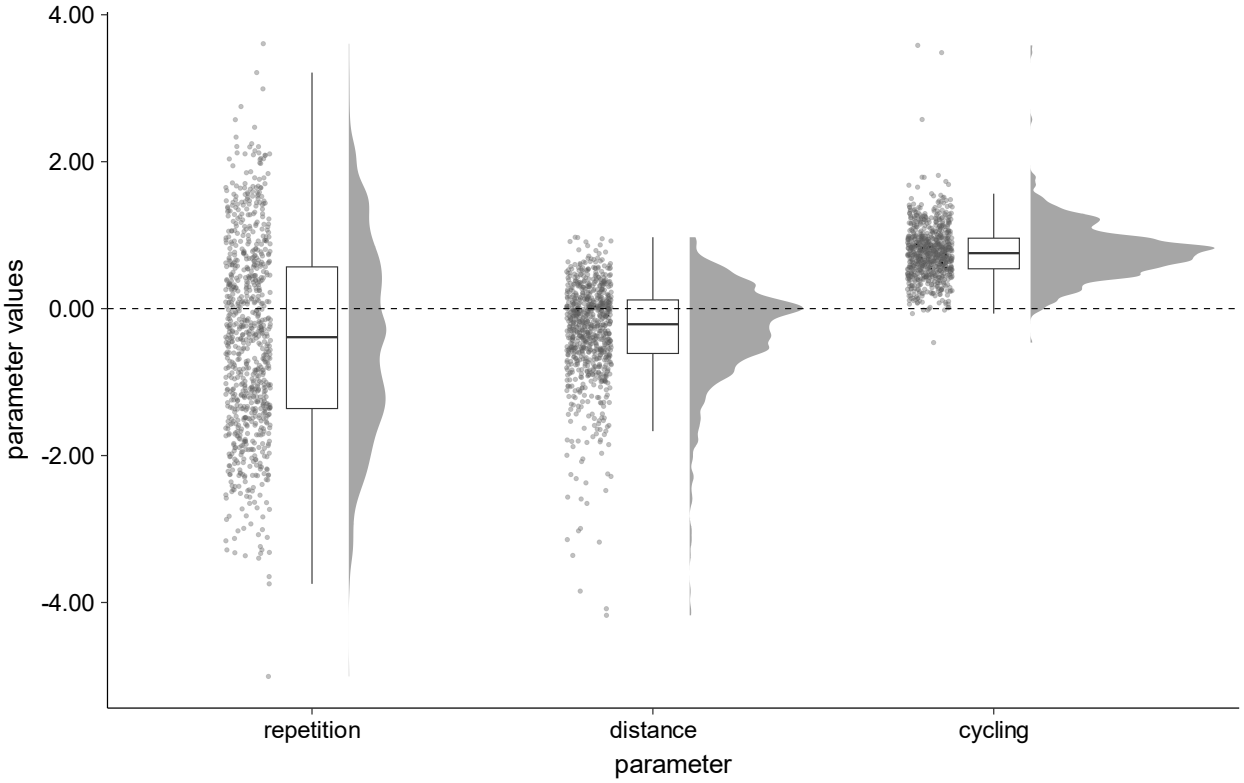
For the analysis of parameter estimates, we only used the results of the three-parameter model with the logarithmic scaling of the gap to the last occurrence of a number as all metrics of model fit as well as model-driven simulations indicate that this model provided the best approximation to the biases underlying human RNG. First, we investigated the hyperparameters

of the model representing the mean and the standard deviation of the individual parameters. Bayesian credible 95% intervals are provided by the values within square brackets. Participants showed on average a tendency to avoid repetitions (M = -0.41 [-0.53, -0.29], SD = 1.48 [1.38, 1.59]); however, the standard deviation of this parameter was large, much larger than for all other parameters. It appears that despite the general trend to avoid repetitions, some individuals showed an opposite tendency and thus, an increased chance of repetitions. The distance parameter was on average negative (M = -0.32 [-0.36, -0.27], SD = 0.69 [0.65, 0.73], indicating a tendency to choose adjacent pairs (e.g., choosing a 9 after an 8). However, there were some participants that showed the opposite tendency and avoided adjacent numbers. The cycling parameter was positive (M = 0.77 [0.74, 0.79], SD = 0.37 [0.35, 0.40]), indicating a clear tendency to cycle too fast through all available numbers; the probability of choosing a number increased the longer it was not used. For all three parameters, 95 % credible intervals for the hyperparameters representing the means of the individual parameters did not contain the value 0. This finding supports the interpretation that all three parameters can account for a substantial part of human behavior.

We also computed all three parameters for each participant by taking the mean over the posterior simulations of the respective participant. The distributions of the individual repetition, distance, and cycling parameters are shown in Figure 3.

Figure 3

Distribution of the Individual Repetition, Distance, and Cycling Parameter Estimates for the Three-Parameter Model with Logarithmic Scaling of the Gap



Note. The distribution for each parameter is represented in three plots: a density plot (in arbitrary units), a boxplot, and a scatter plot where the variation along the *x*-axis was introduced to allow better representation of the data points. The bold line in the boxplot represents the median. The whiskers of the boxplot are limited to the IQR * 1.50. The dashed line at $y = 0.00$ indicates the absence of a bias.

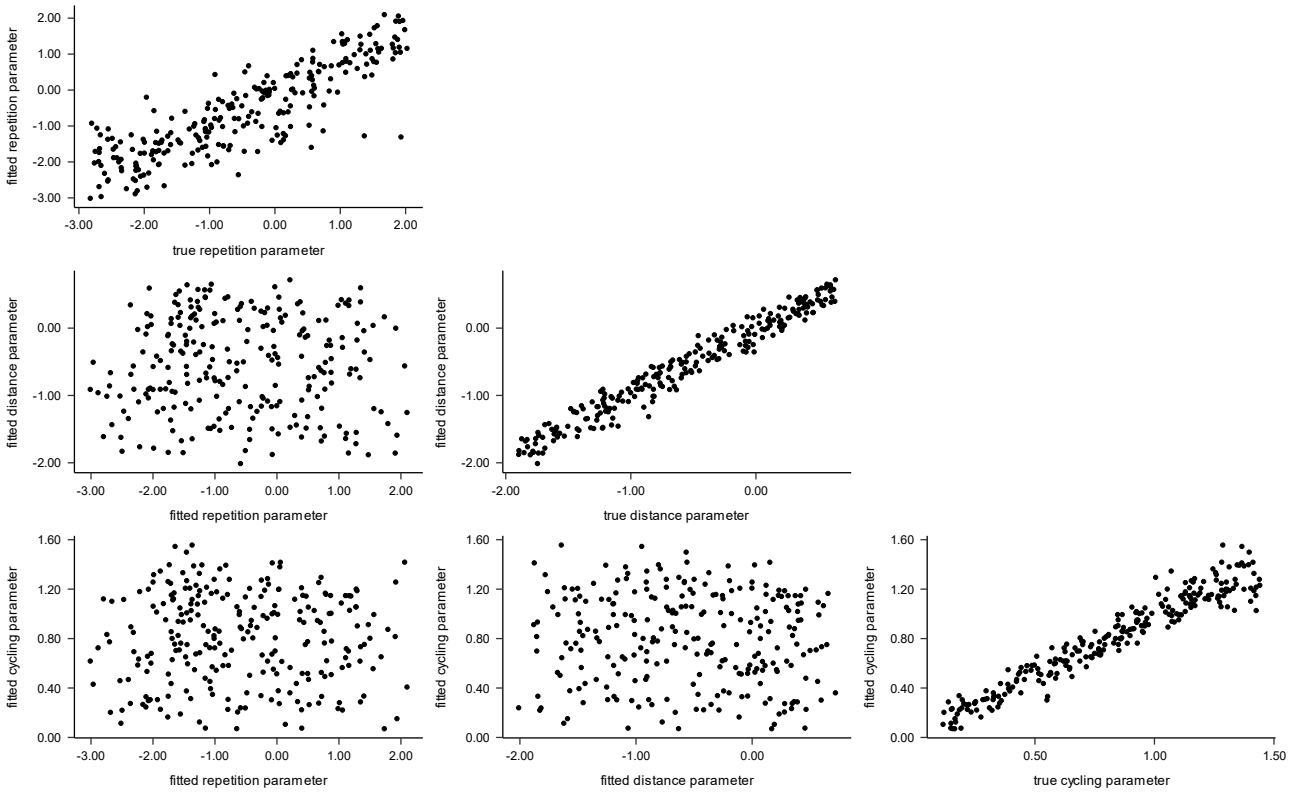
Parameter Recovery and Correlation

Finally, we assessed the reliability of the obtained model estimates from the preceding section in estimating the true parameter values. To this end, we generated 250 number sequences of length 200, comprising numbers 1 to 9, using the three-parameter model with logarithmic gap scaling, to mimic human-generated sequences. The parameter range for the simulation was established by using the 2.5th and 97.5th percentiles of the individual-level parameter estimates from the human-generated sequences as the lower and upper bounds, respectively. Values within this range were simulated using a uniform distribution. The model was then fitted to the

simulated sequences. The fitted parameter values of the simulated sequences were correlated with the true parameter values used for the simulation, avoiding any subjective evaluations. We examined the correlation between the fitted parameter values of various parameters to make sure that they measure different biases in human RNG. The results are shown in Figure 4. It is evident that all three parameters could be retrieved, as evidenced by the robust linear relationship between the true and fitted parameters traced on the scatter plot matrix's diagonal. This means that the model allowed us to reliably gauge the generating parameters of the sequences. We did not detect any apparent correlation between the various fitted parameter values on the left of the diagonal, suggesting that all parameters could be estimated independently.

Figure 4

Parameter Recovery and Parameter Correlations for the Three-Parameter Model with the Logarithmic Scaling of the Gap



Note. The diagonal scatter plots represent the parameter recovery results (correlation between true and fitted parameters). The plots to the left of the diagonal represent correlations between different parameters.

Discussion

We have enhanced Yousif et al.'s (2022) computational model by substituting the side-switching parameter with a distance parameter that is less dependent on procedural details of the employed RNG task. Additionally, we added a cycling parameter to account for the phenomenon that humans tend to cycle through all available numbers too quickly. Our findings indicate that a distance parameter is more effective in explaining human behavior in an RNG task than a side-switching parameter. Adding a cycling parameter improved the model fit further. Simulations using the fitted parameters revealed that according to various measures of randomness, our three-parameter model provided an improved approximation to the characteristics of human-generated

random sequences. Furthermore, we found that model parameters could reliably be recovered in a simulation, and a lack of correlations between the parameters suggested that all model parameters could be estimated reliably and independently.

Replacing the side-switching parameter in the initial Yousif et al. (2022) model with a generalized distance parameter improved the model's fit to human-generated sequences. Our findings are in line with the findings by Towse (1998) who showed that particularly those numbers adjacent or nearby to the previously generated number are selected with increased probability. The distance parameter offers the advantage of being independent of the particular RNG task's format, as it can be tailored precisely to it. In the present study, the sequences generated by humans were obtained in an RNG task that required the selection of numbers within a 3x3 matrix. The utilization of Euclidean distances between numbers within this two-dimensional space unequivocally enhanced the compatibility of the model to the human-generated sequences. This divulges the significance of incorporating parameters of the RNG task into the modelling process.

The inclusion of the cycling parameter appears to be the most beneficial addition to the model proposed by Yousif et al. (2022). This can be deduced from the distribution of parameter estimates of the three-parameter model with the logarithmic scaling of the gap: the hyperparameter representing the mean of all individual cycling parameters is more than two standard deviations above the point of 0 and the 95% Bayesian credible interval of this parameter does not include 0. The repetition and the distance parameters make a considerably smaller contribution to the performance of the three-parameter model. The cycling parameter can account for more complex behavior as it is influenced by an individual's entire prior choice history. In contrast, a repetition, side-switching, or distance parameter only reflects the

immediately preceding choice and therefore cannot capture the same level of behavioral complexity. This interpretation aligns with prior research, as the inclination to quickly cycle through all feasible numbers among human participants has been frequently observed and probed (Ginsburg & Karpiuk, 1994; Peters et al., 2007; Towse & Neil, 1998). Simulations showed that our expanded model is capable of closely approximating human behavior in RNG tasks. However, despite providing a close approximation to the distribution of the phi index, the coupon, and the mean gap score, the model could not fully mimic the distribution of block entropy scores. Nevertheless, it is noteworthy that the approximation was consistently superior to not applying a model, as evidenced by the distribution of randomness measures computed for randomly permuted human-generated sequences.

With the only exception of block entropy, the three-parameter model with a logarithmic scaling of the gap yielded better approximations to human-generated sequences than the two-parameter model. In order to achieve an even better fit to human behavior, further research is necessary. Yousif et al. (2022) reported evidence supporting the suitability of their model by applying repetition and side-switching parameters in various RNG task paradigms. Similar future investigations will likely profit from adding a distance and a cycling parameter.

However, a cycling parameter alone cannot sufficiently account for all human behavior in RNG tasks because on average, the distance parameter assumed negative values, implying participants picked numbers that were close to previously generated ones. This aligns with Towse's (1998) prior research. Participants, on average, tended to avoid repetitions, which previous research has identified as a defining feature of human RNG (Cooper, 2016; Towse, 1998; Towse & Cheshire, 2007). Nonetheless, it is noteworthy that although a general tendency to avoid repetitions was observed, the variability in this parameter was high, with some

participants exhibiting positive parameter values. This finding suggests that some participants do not show an aversion to repetition when other biases like the tendency to cycle through all available numbers too quickly are also taken into account.

When implementing a computational model such as the one used in the present study, it is necessary to formalize all assumptions about human RNG as model equations prior to analysis. This theory-driven approach goes beyond previous research evaluating bias in human RNG tasks. Most studies of human RNG focus on calculating various measures of randomness for the sequences generated by participants. These measures are then often combined using principal component analysis. Researchers have used this approach extensively (Ginsburg & Karpiuk, 1994; Oomens et al., 2015; Peters et al., 2007; Towse & Neil, 1998). However, measures of randomness can only reflect certain statistical patterns demonstrated by humans in RNG tasks; they cannot, by themselves, be interpreted as latent variables underlying human performance in RNG (Oomens et al., 2023). A model-based approach, in contrast, allows to directly and jointly quantify various human biases in RNG tasks, and model parameters are always easily interpreted as specified by the model equations.

The computational model described in this paper can be used in various areas of psychological research that examines RNG performance. For example, it can be used to compare the randomization performance of various clinical groups, such as schizophrenic patients (Hornero et al., 2006; Peters et al., 2007), or to explore the impact of age on RNG performance (Gauvrit et al., 2017; Heuer et al., 2010; Multani et al., 2016). Other applications include testing the impact of auditory distraction (Marsh et al., 2013), manipulating production speed (Towse, 1998), or evaluating the influence of sleep deprivation on RNG performance (Heuer et al., 2005).

In conclusion, we have expanded upon the computational model introduced by Yousif et al. (2022) by incorporating two additional parameters: a distance parameter, a more task-independent variant of the original side-switching parameter, and a cycling parameter, which can account for the propensity of many humans to cycle rapidly through all possible numbers in an RNG task. Including both parameters significantly improved the fit of the model and resulted in a better approximation of human behavior in RNG tasks. The cycling parameter is capable of taking into account not only the impact of the last pick on the selection of the subsequent pick in a sequence, but also the impact of an individual's entire previous selection history. The present promising results suggest that a cycling parameter ought to be integrated into any potential endeavor to assess or simulate human biases in RNG tasks.

Declarations

Funding

This research was conducted at the University of Duesseldorf. No external funding was received for conducting this study.

Conflicts of Interest/Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Consent to Participate

Informed consent was obtained from all individual participants included in the study.

Consent for Publication

All participants provided informed consent regarding publishing their data.

Availability of Data

The datasets generated during and/or analyzed during the current study are available in the OSF repository, https://osf.io/xwzup/?view_only=1052e095327241d280e5602762a66f77.

Code Availability

The code used to analyze the data is available in the OSF repository, https://osf.io/sygdu/?view_only=8c602343aad54fe9a429903d3d553ceb

Open Practices Statement

The data for all experiments are available at https://osf.io/xwzup/?view_only=1052e095327241d280e5602762a66f77.

References

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. A. (2019). Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, *4*, 63. https://doi.org/10.12688/wellcomeopenres.15191.1

Angelike, T. (2022). *randfindR: Analysis of randomness in human generated sequences* [Computer software]. https://github.com/TImA97/randfindR

Angelike, T., & Musch, J. (2023). *A comparative evaluation of measures to assess randomness in human-generated sequences* [manuscript submitted for publication]

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown* [Computer software]. https://github.com/crsh/papaja

Bocharov, A., Freedman, M., Kemp, E., Roetteler, M., & Svore, K. M. (2020). *Predicting human-generated bitstreams using classical and quantum models*. arXiv. https://doi.org/10.48550/ARXIV.2004.04671

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221–235). Russell Sage Foundation. https://www.russellsage.org/publications/handbook-research-synthesis-and-meta-analysis-second-edition

Capone, F., Capone, G., Ranieri, F., Di Pino, G., Oricchio, G., & Di Lazzaro, V. (2014). The effect of practice on random number generation task: A transcranial direct current stimulation study. *Neurobiology of Learning and Memory*, *114*, 51–57. https://doi.org/10.1016/j.nlm.2014.04.013

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Cooper, R. P. (2016). Executive functions and the generation of "random" sequential responses: A computational account. *Journal of Mathematical Psychology*, *73*, 153–168. https://doi.org/10.1016/j.jmp.2016.06.002

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, *104*, 301–318. https://doi.org/10.1037/0033-295X.104.2.301

Figurska, M., Stańczyk, M., & Kulesza, K. (2008). Humans cannot consciously generate random numbers sequences: Polemic study. *Medical Hypotheses*, *70*(1), 182–185. https://doi.org/10.1016/j.mehy.2007.06.038

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General*, *133*(1), 101–135. https://doi.org/10.1037/0096-3445.133.1.101

Gauvrit, N., Singmann, H., Soler-Toscano, F., & Zenil, H. (2016). Algorithmic complexity for psychology: A user-friendly implementation of the coding theorem method. *Behavior Research Methods*, *48*(1), 314–329. https://doi.org/10.3758/s13428-015-0574-3

Gauvrit, N., Zenil, H., Soler-Toscano, F., Delahaye, J.-P., & Brugger, P. (2017). Human behavioral complexity peaks at age 25. *PLOS Computational Biology*, *13*(4), e1005408. https://doi.org/10.1371/journal.pcbi.1005408

Ginsburg, N., & Karpiuk, P. (1994). Random generation: Analysis of the responses. *Perceptual and Motor Skills*, *79*(3), 1059–1067. https://doi.org/10.2466/pms.1994.79.3.1059

Heuer, H., Janczyk, M., & Kunde, W. (2010). Random noun generation in younger and older adults. *Quarterly Journal of Experimental Psychology*, *63*(3), 465–478. https://doi.org/10.1080/17470210902974138

Heuer, H., Kohlisch, O., & Klein, W. (2005). The effects of total sleep deprivation on the generation of random sequences of key-presses, numbers and nouns. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *58*(2), 275–307. https://doi.org/10.1080/02724980343000855

Hornero, R., Abásolo, D., Jimeno, N., Sánchez, C. I., Poza, J., & Aboy, M. (2006). Variability, regularity, and complexity of time series generated by schizophrenic patients and control subjects. *IEEE Transactions on Biomedical Engineering*, *53*(2), 210–218. https://doi.org/10.1109/TBME.2005.862547

Kassambara, A. (2020). *ggpubr: "ggplot2" based publication ready plots* [Computer software]. https://CRAN.R-project.org/package=ggpubr

Kay, M. (2023). *ggdist: Visualizations of distributions and uncertainty* [Computer software]. https://doi.org/10.5281/zenodo.3879620

Kee, Y. H., Chaturvedi, I., Wang, C. K. J., & Chen, L. H. (2013). The power of now: Brief mindfulness induction led to increased randomness of clicking sequence. *Motor Control*, *17*(3), 238–255. https://doi.org/10.1123/mcj.17.3.238

Maes, J. H. R., Eling, P. A. T. M., Reelick, M. F., & Kessels, R. P. C. (2011). Assessing executive functioning: On the validity, reliability, and sensitivity of a click/point random number generation task in healthy adults and patients with cognitive decline. *Journal of Clinical and Experimental Neuropsychology*, *33*(3), 366–378. https://doi.org/10.1080/13803395.2010.524149

Marsh, J. E., Sörqvist, P., Halin, N., Nöstl, A., & Jones, D. M. (2013). Auditory distraction compromises random generation: Falling back into old habits? *Experimental Psychology*, *60*(4), 279–292. https://doi.org/10.1027/1618-3169/a000198

Moore, D. G., Valentini, G., Walker, S. I., & Levin, M. (2018). Inform: Efficient information-theoretic analysis of collective behaviors. *Frontiers in Robotics and AI*, *5*. https://doi.org/10.3389/frobt.2018.00060

Morey, R. D., & Rouder, J. N. (2022). *Bayesfactor: Computation of Bayes Factors for common designs* [Computer software]. https://CRAN.R-project.org/package=BayesFactor

Multani, N., Rudzicz, F., Wong, W. Y. S., Namasivayam, A. K., & van Lieshout, P. (2016). Random item generation is affected by age. *Journal of Speech, Language, and Hearing Research*, *59*(5), 1172–1178.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review*, *109*(2), 330–357. https://doi.org/10.1037/0033-295X.109.2.330

Oomens, W., Maes, J. H. R., Hasselman, F., & Egger, J. I. M. (2015). A time series approach to random number generation: Using recurrence quantification analysis to capture executive behavior. *Frontiers in Human Neuroscience*, *9*(JUNE). https://doi.org/10.3389/fnhum.2015.00319

Oomens, W., Maes, J. H. R., Hasselman, F., & Egger, J. I. M. (2021). RandseqR: An R package for describing performance on the random number generation task. *Frontiers in Psychology*, *12*, 629012. https://doi.org/10.3389/fpsyg.2021.629012

Oomens, W., Maes, J. H. R., Hasselman, F., & Egger, J. I. M. (2023). A time-series perspective on executive functioning: The benefits of a dynamic approach to random number

generation. *International Journal of Methods in Psychiatric Research*, *32*(2), e1945. https://doi.org/10.1002/mpr.1945

Pedersen, T. L. (2022). *patchwork: The composer of plots* [Computer software]. https://CRAN.R-project.org/package=patchwork

Peters, M., Giesbrecht, T., Jelicic, M., & Merckelbach, H. (2007). The random number generation task: Psychometric properties and normative data of an executive function task in a mixed sample. *Journal of the International Neuropsychological Society*, *13*(4), 626–634. https://doi.org/10.1017/S1355617707070786

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Schulz, M.-A., Baier, S., Timmermann, B., Bzdok, D., & Witt, K. (2021). A cognitive fingerprint in human random number generation. *Scientific Reports*, *11*(1), 20217. https://doi.org/10.1038/s41598-021-98315-y

Sexton, N. J., & Cooper, R. P. (2014). An architecturally constrained model of random number generation and its application to modeling the effect of generation rate. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00670

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shteingart, H., & Loewenstein, Y. (2016). Heterogeneous suppression of sequential effects in random sequence generation, but not in operant learning. *PLOS ONE*, *11*(8), e0157643. https://doi.org/10.1371/journal.pone.0157643

Stan Development Team. (2022). *RStan: The R interface to Stan* [Computer software]. https://mc-stan.org/

Tiedemann, F. (2022). *gghalves: Compose half-half plots using your favourite geoms* (0.1.4) [Computer software]. https://cran.r-project.org/web/packages/gghalves/index.html

Torchiano, M. (2020). *Effsize: Efficient effect size computation* [Computer software]. https://doi.org/10.5281/zenodo.1480624

Towse, J. N. (1998). On random generation and the central executive of working memory. *British Journal of Psychology*, *89*(1), 77–101. https://doi.org/10.1111/j.2044-8295.1998.tb02674.x

Towse, J. N., & Cheshire, A. (2007). Random number generation and working memory. *European Journal of Cognitive Psychology*, *19*(3), 374–394. https://doi.org/10.1080/09541440600764570

Towse, J. N., & Neil, D. (1998). Analyzing human random generation behavior: A review of methods used and a computer program for describing performance. *Behavior Research Methods, Instruments, & Computers*, *30*(4), 583–591. https://doi.org/10.3758/BF03209475

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110. https://doi.org/10.1037/h0031322

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Yousif, S. R., McDougle, S. D., & Rutledge, R. B. (2022). A task-general model of human randomization. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*. https://escholarship.org/uc/item/978107w8

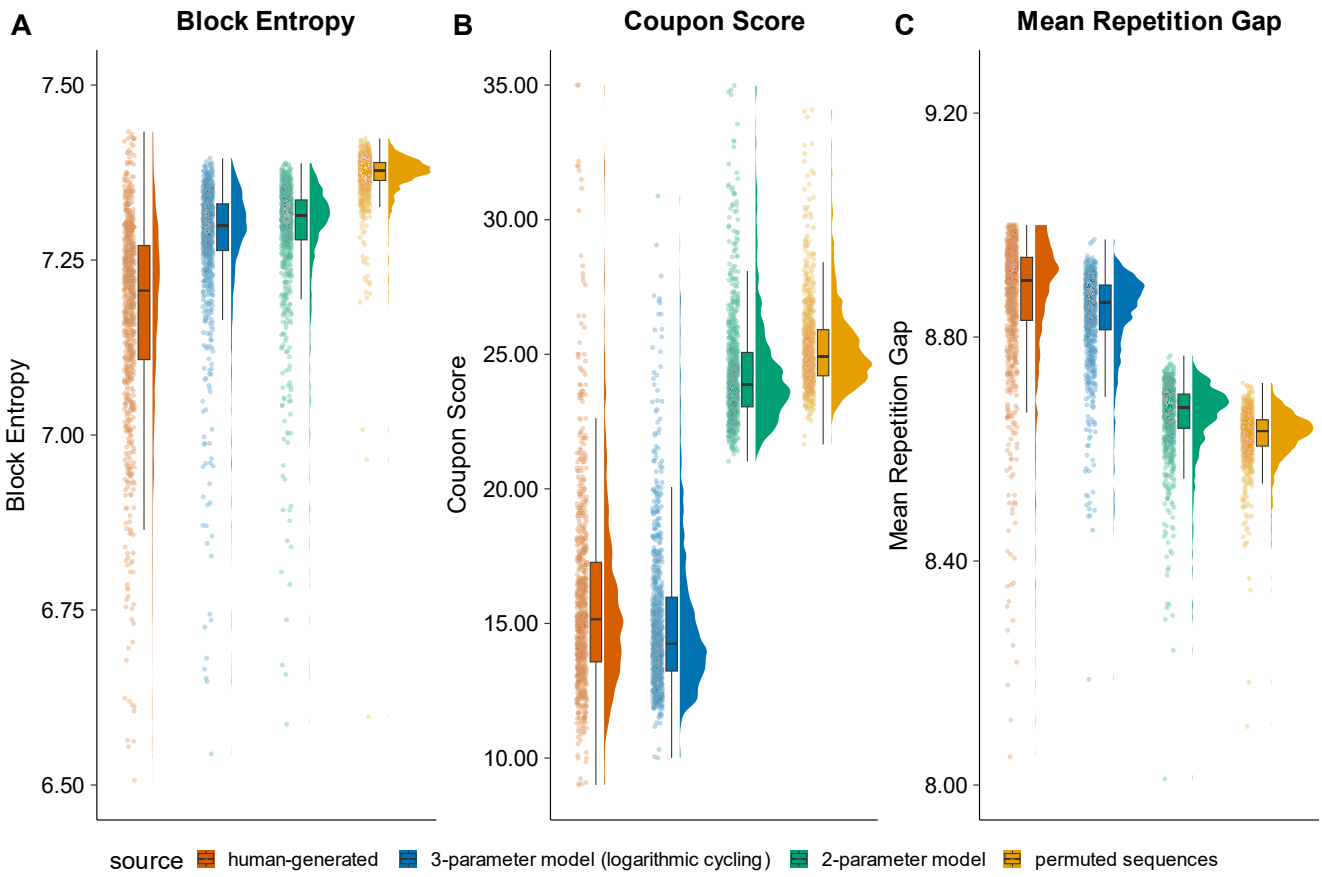
Appendix

Simulation of Model-Driven Data (Continued)

Figure A1 shows the distributions of three measures of randomness (block entropy for blocks of size 3, coupon score, and mean gap score) for data that were generated by humans, by the three-parameter model with a logarithmic scaling of the gap, by the two-parameter model with a distance instead of a side-switching parameter, and by randomly permuting human-generated sequences. As can be seen, the distribution of block entropy scores is least well approximated by the three- and the two-parameter model, though the approximation appears to be an improvement to the distribution of block entropy scores for randomly permuted sequences (human-generated: $M = 7.12$, $SD = 0.38$; three-parameter model: $M = 7.27$, $SD = 0.18$; two-parameter model: $M = 7.28$, $SD = 0.15$; permuted: $M = 7.37$, $SD = 0.04$). The three-parameter model also provided the closest approximation to the distribution of the coupon score (human-generated: $M = 16.00$, $SD = 4.15$, three-parameter model: $M = 15.36$, $SD = 6.25$, two-parameter model: $M = 24.65$, $SD = 4.22$, permuted: $M = 25.47$, $SD = 2.83$). The same was true for the mean repetition gap score (human-generated: $M = 8.86$, $SD = 0.22$, three-parameter model: $M = 8.83$, $SD = 0.18$, two-parameter model: $M = 8.65$, $SD = 0.13$, permuted: $M = 8.62$, $SD = 0.06$).

Figure A1

Distribution of Randomness Measures for Human-Generated, Model-Generated, and Randomly Permuted Human-Generated Sequences



Note. This plot shows the distribution of the block entropy for blocks of size 3, the coupon score, and the mean gap score computed over different sources of sequences: sequences generated by humans in a RNG task, sequences generated by the three-parameter model with the logarithmic scaling of the gap to the last occurrence of a number, the two-parameter model with a distance instead of a side-switching parameter, and sequences that were random permutations of the original-human-generated sequences. The distribution for each source is represented in three plots: a density plot (in arbitrary units), a boxplot, and a scatter plot where the variation along the *x*-axis was introduced to allow better representation of the data points. The bold line in the boxplot represents the median. The whiskers of the boxplot are limited to the IQR * 1.50. Block entropy was computed over blocks of size 3 as these block sizes were found to discriminate best between human-generated and random sequences in Angelike and Musch (2023). Long tails of

the distributions had to be excluded for the sake of clarity of visualization (there were 40 values below 6.50 and above 7.50 for block entropy, 22 values above 35 for the coupon score, and eight values below 8 and above 9.25 for the mean gap score).

Regarding block entropy scores, both models produced sequences that more closely mimicked human-generated sequences than permuted sequences (three-parameter model: correct classification rate = 71.61% [68.79%, 74.49%], two-parameter model: correct classification rate = 75.03% [72.32%, 77.76%], permuted: correct classification rate = 92.72% [91.00%, 94.24%]). However, due to the overlapping confidence intervals, the data do not show a clear superiority of the three-parameter model over the two-parameter model. Regarding the coupon score, the approximation of the human-generated sequences through the three-parameter model was significantly better than the approximation with the two-parameter model (three-parameter model: correct classification rate = 56.96% [53.88%, 60.10%], two-parameter model: correct classification rate = 95.44% [94.10%, 96.72%], permuted: correct classification rate = 96.40% [95.31%, 97.63%]). A similar result could be obtained for the mean gap score (three-parameter model: correct classification rate = 55.31% [52.29%, 58.35%], two-parameter model: correct classification rate = 93.17% [91.65%, 94.79%], permuted: correct classification rate = 94.79% [93.36%, 96.17%]). Taken together, our results indicate that the three-parameter model produced sequences for which the distribution of the randomness measures under investigation mimicked the distribution of the randomness measures for human-generated sequences better than both, the sequences produced with the two-parameter model with a distance parameter and the sequences generated by randomly permuting human-generated sequences.