

# THE **SEO** **COMPRESSION** MYTH

What Analysis of **42,00 Websites**  
Revealed About Content Compression  
& Ranking Correlations



# CONTENTS

---

Foreword .....	3
<b>Introduction</b> .....	4
<b>Our Hypothesis</b> .....	5
<b>Methodology</b> .....	5
Our Seed List .....	6
Content Compression .....	8
Data Aggregation .....	8
<b>Results &amp; Findings</b> .....	9



# Foreword

Within the SEO industry we are fortunate that many ideas and theories are shared which can help develop everyone's understanding of complex topics and result in a good volume of interesting debate.

As you engage with the SEO community you will find no shortage of discussion surrounding what does and doesn't work when it comes to improving the organic rankings of a given website. This is perhaps one of the best things about working in this space.

However, there is also the tendency within the industry to accept ideas, especially more complex and/or advanced technical ones, without extensive testing and individual experimentation.

For over a decade at **Reboot Online** we have prioritised experimentation, resulting in an

extensive collection of internal and publicly available **SEO experiments** and research pieces which have put many of the most common SEO myths, theories and misconceptions to the test.

We continue to experiment to learn more about the inner workings of the most popular search engines and AI platforms, and to publish much of our research in the hope that it can move the industry forwards and generate even more insights and debate.

If you would like to discuss our experiments further, and hear a bit about how we might be able to help you accelerate the SEO growth of your website, please **get in touch**.

**Oliver Sissons**  
Search Director,  
Reboot Online



We ran large scale SEO analysis on **42,000** webpages across multiple markets (UK, US, FR, AU, CA, DE) to find out if content compression ratios correlated with stronger organic rankings.

Our analysis included crawling **54,000 websites** in total across various industries including:

- Finance
- eCommerce
- Travel
- Gaming
- Technology
- Healthcare
- Education
- Real Estate
- Retail
- Energy
- Food & Beverages
- Logistics & Transport
- Automotive
- Media & Entertainment

*After removing webpages which returned errors and duplicates, we were left with 42,000 unique webpages to use in our study.*



# Introduction



We know that Google is using many different signals to help determine where a given webpage should rank in the organic search results, and one long-held belief by some SEOs is that content compression could be one such signal.

Several SEOs have shared the theory that if a piece of content can be highly compressed, that is to say that many repetitive/redundant words and phrases can be removed to reduce the overall size of the HTML file, this could suggest that the content is of a lower quality (otherwise it would contain a greater ratio of unique words and phrases, meaning that the HTML file couldn't be compressed as much).

**On the surface, this theory is a logical one.**

Indexing the entire internet and ranking results for billions of search queries each and every day takes a staggering amount of resources, not to mention the environmental impact associated with this enormous task.

The idea that looking at things like compression ratios to help determine not only content quality, but also which pages should or shouldn't be crawled (and/or recrawled) and indexed to begin with, is an intuitive one.

Doing so could help search engines allocate valuable resources more effectively, and reduce energy/resources being wasted on lower quality pages and content.

If true, this practice would also mean that we could automate content quality checks by looking at the compression ratios of our own content relative to the ratios of the webpages ranking top for our target keywords.

When doing so, if we found that our pages could be compressed to a greater extent than the top ranking ones, this could suggest that Google would find the same and deem our content lower quality than the competitors. If this were the case, it could greatly impact our ability to rank for our target keywords.

Given the clear opportunity presented to develop a bespoke tool to help with these kinds of automated content quality checks and competitor comparisons, if the content compression theory were proved to be true, and what this would mean for our largest-scale SEO campaigns, we were keen to put the theory to the test.

That is why we developed this study and carried out this analysis, and the results (which we will outline below) were surprising to say the least.

# Our Hypothesis

Once we decided that we wanted to put the content compression theory to the test, we first needed to confirm the hypothesis that we were looking to prove or disprove.



**After some internal discussions with our data team, we arrived at the following hypothesis:**

“If compression ratios are being used as an indicator of content quality by the major search engines, we should find that webpages ranking in the top 10 organic positions for a range of keywords can be compressed less (on average) than the webpages ranking in the bottom 20 positions for those same search terms.”

**Paul Lapham**

*Senior Data Insights Manager,  
Reboot Online*



# Methodology

Since in this study we were interested in determining if and how major search engines like Google could be using compression ratios in their algorithms to infer content quality and influence rankings, we knew that our analysis should involve looking for a correlation between compression ratios and search engine position.

From the outset, we also knew that we would need to look at a large number of webpages and search results when carrying out this study.

Finding that strong rankings for a small sample of keywords in a single industry correlated with a smaller compression ratio wouldn't be nearly enough for us to confidently determine if content compression ratios are being used at a large scale to influence the organic search results.

So, we devised a methodology that would look at compression ratios and any correlation between these and search engine rankings across multiple markets and industries.

The first step in carrying out this study was building out our seed list of keywords and webpages to analyse.



# Seed List

To build our seed list, we came up with **15 different industries** and looked at **20 different keywords within each**, across **6 different markets**.

Our analysis spanned the below markets:



The keywords for each industry looked like this: “[keyword] [country]”.

For example, one keyword used within the gaming industry was “VR France”.

The full list of industries and keywords can be found to the right:

Industry	Keywords	Industry	Keywords
<b>Finance</b> 	Investment, Cryptocurrency, Banking, Wealth Management, Stock Market, Risk Analysis, Financial Planning, Loans, Credit, Accounting, Auditing, Taxes, Assets, ROI, Bonds, Insurance, Portfolio Management, Equity, Fintech, Capital	<b>Retail</b> 	Point of Sale, Merchandising, Inventory Management, Brick-and-Mortar, Loyalty Programs, Retail Analytics, Supply Chain, Store Layout, Discounts, Retail Trends, Consumer Behavior, Personalization, Omnichannel, Visual Merchandising, Product Launch, Retail Marketing, Foot Traffic, Customer Experience, Seasonal Sales, Upselling
<b>eCommerce</b> 	Online Shopping, Dropshipping, Marketplace, Product Listing, SEO, Digital Marketing, Payment Gateway, Fulfillment, Customer Retention, Conversion Rate, Cart Abandonment, Multichannel Selling, Logistics, Mobile Commerce, Inventory Management, AI Personalization, Reviews, Affiliate Marketing, Analytics, Omnichannel	<b>Energy</b> 	Renewable Energy, Solar Power, Wind Energy, Energy Storage, Smart Grids, Fossil Fuels, Energy Efficiency, Carbon Neutrality, Oil & Gas, Clean Technology, Nuclear Energy, Geothermal Energy, Energy Trading, Biofuels, Energy Policy, Sustainability, EV Charging, Battery Technology, Energy Audits, Grid Modernization
<b>Travel</b> 	Tourism, Booking, Flights, Hotels, Travel Insurance, Itinerary, Cruises, Travel Agencies, Adventure Tours, Group Travel, Backpacking, Luxury Travel, Eco-tourism, Vacation Rentals, Destination Marketing, Budget Travel, Sightseeing, Online Travel Agencies, Loyalty Programs, Travel Tech	<b>Fashion</b> 	Sustainable Fashion, Fast Fashion, Brand Identity, Luxury Fashion, Textile Innovation, Fashion Shows, Trend Forecasting, E-commerce, Ethical Sourcing, Digital Runways, Circular Fashion, Streetwear, Customization, Influencer Marketing, Apparel Manufacturing, Fashion Tech, Retail Analytics, Omnichannel Selling, Fashion Startups, Fabric Design
<b>Gaming</b> 	eSports, Game Development, Streaming, Virtual Reality, Mobile Gaming, In-App Purchases, Multiplayer, Game Engines, Indie Games, Console Gaming, Cloud Gaming, NFTs, Game Monetization, Player Engagement, DLCs (Downloadable Content), AI in Gaming, Augmented Reality, Leaderboards, Game Studios, Subscription Gaming	<b>Food &amp; Beverages</b> 	Farm-to-Table, Supply Chain, Food Safety, Organic Foods, Packaging, Delivery Apps, Food Trucks, Sustainability, Culinary Trends, Franchising, Menu Innovation, Beverage Technology, Seasonal Menus, Restaurant Management, Plant-Based Foods, Nutritional Labeling, Food Waste Management, Branding, CPG (Consumer Packaged Goods), Subscription Boxes
<b>Technology</b> 	AI, Machine Learning, IoT, Cybersecurity, Cloud Computing, SaaS, Blockchain, Big Data, Robotics, AR/VR, Software Development, 5G, Digital Transformation, Edge Computing, Automation, Smart Devices, IT Infrastructure, DevOps, Data Science, Green Tech	<b>Logistics &amp; Transport</b> 	Supply Chain, Freight, Last-Mile Delivery, Fleet Management, Warehousing, Logistics Tech, Route Optimization, Shipping, Customs Clearance, Import/Export, Air Freight, Containerization, Reverse Logistics, Packaging Solutions, Cold Chain, Delivery Drones, Supply Chain Visibility, Freight Forwarding, Cargo Insurance, Green Logistics
<b>Healthcare</b> 	Telemedicine, Patient Care, Pharmaceuticals, Medical Devices, Healthcare IT, Diagnostics, Preventive Care, Biotech, Electronic Health Records, Medical Billing, Wellness Programs, Hospital Management, AI in Healthcare, Remote Monitoring, Clinical Trials, Health Insurance, Public Health, Health Tech, Rehabilitation, Nutrition	<b>Automotive</b> 	Electric Vehicles, Autonomous Driving, Connected Cars, Vehicle Maintenance, Auto Parts, Dealerships, Car Rentals, Ride Sharing, Hydrogen Cars, Automotive Tech, EV Charging, Luxury Vehicles, Safety Features, Vehicle Financing, Aftermarket, Telematics, Fleet Management, Auto Insurance, Manufacturing, Regulatory Compliance
<b>Education</b> 	E-learning, EdTech, Curriculum Design, Student Engagement, Online Courses, Assessment Tools, STEM Education, Certification Programs, Digital Classrooms, Gamification, Lifelong Learning, Teacher Training, Learning Management Systems, Skill Development, Test Prep, Adaptive Learning, Academic Research, Virtual Learning Environments, Blended Learning, College Admissions	<b>Media &amp; Entertainment</b> 	Content Creation, OTT Platforms, Streaming, Film Production, Music Licensing, Influencer Marketing, Ad Tech, Copyright Management, Virtual Events, Gaming, Animation, Sports Media, Podcasting, Digital Publishing, Social Media, Live Concerts, Media Analytics, AI in Media, Event Management, Cross-Platform Integration
<b>Real Estate</b> 	Property Listings, Mortgage, Rental Market, Real Estate Investment, Home Appraisals, Commercial Property, Real Estate Agents, Virtual Tours, Zoning Laws, Property Management, Housing Market Trends, Home Staging, Land Development, Foreclosures, Luxury Real Estate, Open Houses, Leasing, Real Estate Tech, Construction, Sustainability in Real Estate		



For each keyword, we queried the top 100 organic search results and captured the content found on the pages ranking in the top 10 positions and the bottom 20.

For example, we captured content for the results ranking in positions 1 to 10 and in positions 80 to 99.

When less than 100 results were displayed for a keyword, the second/bottom range was lower, i.e. we captured the content of those webpages ranking in positions 60 to 79 (for example).

For each page captured, the whole text of the page was obtained and cleaned. We did things like removing whitespaces, deleting leftover javascript/HTML, and filtering out other non-relevant artifacts.

The position of each page was also recorded, so we could group them into top 10 or bottom 20 results and reference the organic position of each page later on in our analysis.

Irrelevant results were also removed, including pages where the text suggested errors were present (i.e. "404 error, please try again"), or when functionality on the page was blocking us from getting the actual content (e.g. "Please enable cookies to access this page"). These naturally will have had a low compression ratio and were not relevant to the study.

**In all, we collected content from 54,000 websites.**

Once we had filtered out the irrelevant ones and removed duplicates, we were left with 42,000 unique pages for our study.



# Content Compression

## What is Content Compression?

For those not familiar with the concept of compression, it involves removing redundant information from data (in this case, content).

### Take the following string for example:

AAAABBC

An example of compression for this string would be the following:

4A2B1C



In the original string we have 4 letter As, 2 letter Bs, and 1 letter C, so we can remove the duplicate letters and instead simply list how many of each letter is present in the string.

Doing this reduces the length of the stored string, as there are no longer duplicates present, meaning the file size will be reduced slightly.

This encoded data can then easily be decoded and transformed back into the original.

The more redundant and duplicate information contained in the data (in our case this would mean more repeated phrases and/or keywords), the more compressed the content will become.

## Next, we looked at the compression ratio for each of our 42,000 webpages.

The compression ratio was calculated for each page by first calculating the size of the original text by looking at the size of the UTF-8 encoded text in bytes.

We then used the Python library GZIP to compress the text found on each page in our sample, whose size was then measured in the same way as the original text.

Finally, the compression ratio was calculated by dividing the original text size by the compressed text size.

A webpage that could be compressed more would have a larger compression ratio, and a webpage that could be compressed less would have a smaller ratio.

Looking back on our hypothesis, we would expect to find that the webpages which had a larger compression ratio would rank less strongly in the organic search results than those which had a smaller compression ratio.

# Data Aggregation

The next stage in our study was aggregating the data to check the difference in compression ratios across rankings and industries.

Our data aggregation was done by taking the average of each group.

For each aggregation, each value (cell) had a minimum 900 samples (i.e. the average ratio for the top 10 search results for technology related searches had a sample of 941 pages).

The average compression ratio was also compared by organic position, i.e. the average ratio for all pages in position 1, then 2, then 3 etc. of a Google search were plotted.





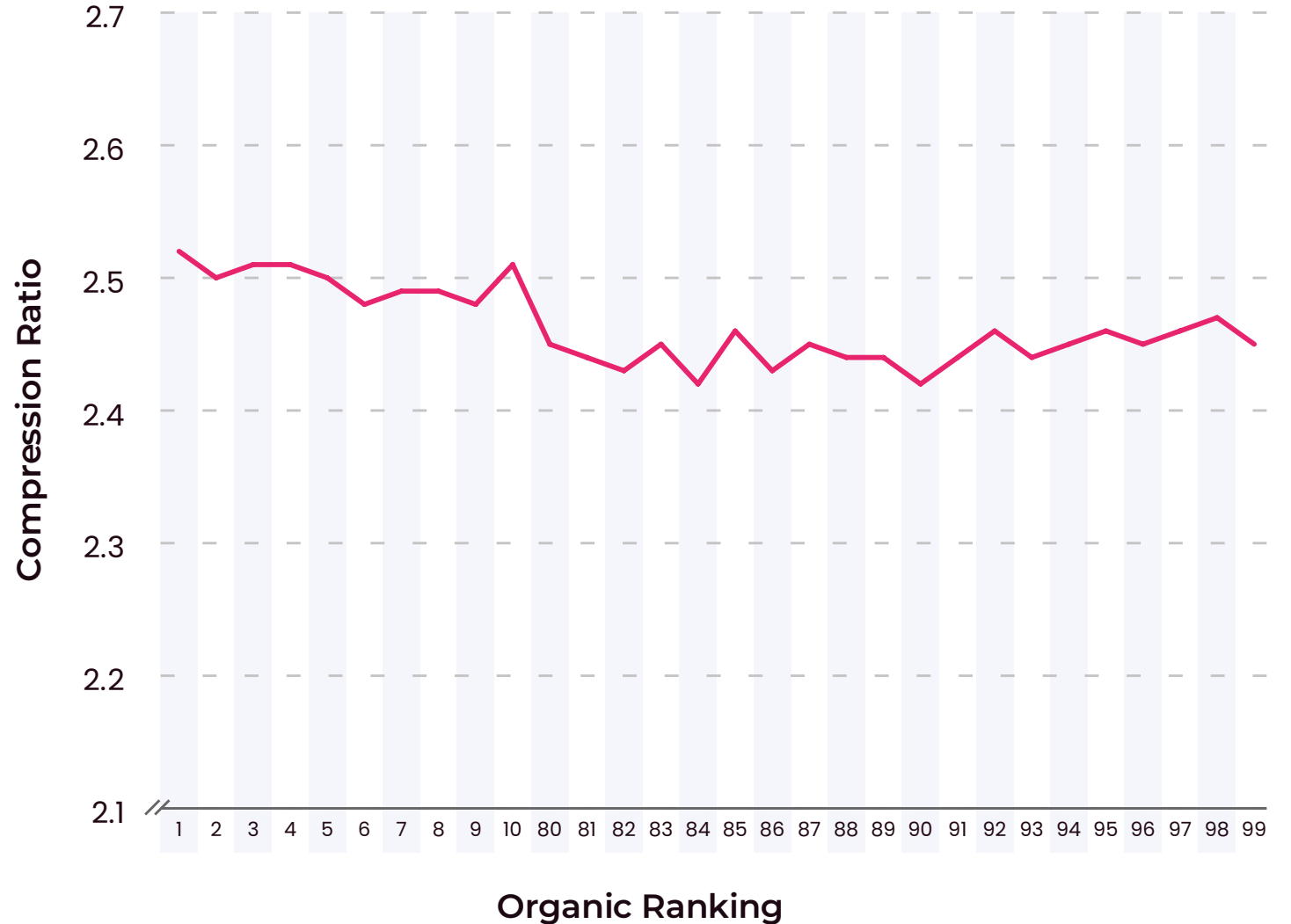
# Results & Findings

Position Group	Compression Ratio
Top 10	2.4982
Bottom	2.443052

Overall, our study found that **compression ratios were actually higher for pages in the top 10 results**, by approximately 2%.

When plotting the compression ratio by organic position, **the ratio appears to decrease the further down the organic results you go**:

## Compression Ratio By Organic Ranking Positions
















This is the **opposite of what we would expect** if our hypothesis were true.













In short, our analysis found **no correlation between compression ratios and organic rankings.**

Since most SEOs accept that Google ranks higher quality content stronger in the organic search results, **this suggests that compression ratios are not a good indicator of content quality.**

We found across all but one industry that the above results remained consistent, meaning **across nearly all industries the webpages ranking in the top 10 organic positions had greater compression ratios than those ranking in the bottom 20 positions:**

Industry	Position Group	Compression Ratio
 Automotive	Top 10	2.462066
 Automotive	Bottom	2.407325
 eCommerce	Top 10	2.593797
 eCommerce	Bottom	2.486386
 Education	Top 10	2.543528
 Education	Bottom	2.50852
 Energy	Top 10	2.414606
 Energy	Bottom	2.427682
 Fashion	Top 10	2.499793
 Fashion	Bottom	2.439908
 Finance	Top 10	2.479606
 Finance	Bottom	2.47359
 Food & Beverage	Top 10	2.483396
 Food & Beverage	Bottom	2.385314
 Gaming	Top 10	2.403172
 Gaming	Bottom	2.373186
 Healthcare	Top 10	2.504533
 Healthcare	Bottom	2.463244
 Logistics & Transport	Top 10	2.443358
 Logistics & Transport	Bottom	2.428211
 Media & Entertainment	Top 10	2.556497
 Media & Entertainment	Bottom	2.431408
 Real Estate	Top 10	2.539945
 Real Estate	Bottom	2.442678
 Retail	Top 10	2.584255
 Retail	Bottom	2.44611
 Technology	Top 10	2.509614
 Technology	Bottom	2.48882
 Travel	Top 10	2.456233
 Travel	Bottom	2.444199

The above trend was also consistent across each market we looked at, with the compression ratios of the webpages ranking in the top 10 organic positions remaining higher than those ranking in the bottom 20 across every market:

Country	Position Group	Compression Ratio
 Australia	Top 10	2.491425
 Australia	Bottom	2.472892
 Canada	Top 10	2.485222
 Canada	Bottom	2.446446
 France	Top 10	2.489321
 France	Bottom	2.415831
 Germany	Top 10	2.502061
 Germany	Bottom	2.451841
 UK	Top 10	2.511284
 UK	Bottom	2.425959
 USA	Top 10	2.509045
 USA	Bottom	2.446743

While this study did not prove our initial hypothesis, it was very interesting to find that actually the opposite of what you would think would happen was found in our dataset.

This serves to highlight the need to run your own tests and experimentations before adopting any widely shared SEO theory.

To read about some of other SEO experiments [click here](#), and please do **get in touch** if you would like to discuss this study (or any of our other ones) and how we can help you drive more SEO growth for your brand.





# REBOOT

 [www.rebootonline.com](http://www.rebootonline.com)

 [hello@rebootonline.com](mailto:hello@rebootonline.com)

 0203 397 1948