項目反応理論を用いたベンチマークからの選択的項目抽出

劔持壮吾　松崎拓也
東京理科大学 理学研究科 応用数学専攻
1425513@ed.tus.ac.jp, matuzaki@rs.tus.ac.jp

概要

本研究では，大規模言語モデルの評価の効率化を目的とし，数学的推論ベンチマークである GSM8K に対して項目反応理論を用いた分析を行った．具体的には，各項目の困難度および識別力に基づきデータセットの項目の抽出を行い，抽出前後におけるモデルの能力パラメータ推定の精度について，ベンチマーク内での順位の安定性を尺度として調査した．分析の結果，困難度に基づく項目抽出では高性能モデルに適した項目が選択される一方，識別力に基づく抽出では全項目の 10% 程度のデータ量であっても，全データを用いた場合と高い順位相関を持つ推定結果が得られることがわかった．

1　はじめに

近年の大規模言語モデル（LLM）の発展に伴い，モデル間で性能を比較することの重要性が増している．特に，各モデルに対してベンチマークと呼ばれる一連のタスク（項目）群を解かせ，正答率などの評価指標により，性能を比較する手法が広く用いられている．

一方で，ベンチマーク内の項目数は数千から数万にまで及ぶことが多く，一度の評価にかかる計算コストが増大している．また，項目数の膨大さゆえに，個別の項目に対する人間による精査が困難となっており，ラベルミスなどの項目の整合性の確認や，個々の項目に対するモデルの挙動の詳細な分析などが難しくなっている．

以上の背景から，評価精度を維持しながらより少ない項目数で効率的な評価を行う手法の開発が望まれる．これを受け，本研究では，個々の項目の特性とモデルの能力を独立に推定可能な**項目反応理論 (IRT)** を用いて，ベンチマークの項目数の削減を試みる．具体的には，各項目の難しさ（**困難度**），およびモデル間の能力の差を見分ける力（**識別力**）に基づき項目を選択的に抽出し，抽出前後での IRT によるモデルの能力値（**潜在能力**）に基づく順位の安定性について，相関係数などの観点から評価を行う．そして，どの抽出方法でどの程度の項目数であれば，元のベンチマークと高い順位一貫性を保てるかを検証する．

本稿では，まず関連研究を概観した後，IRT の基礎的な概念と本研究で用いた手法について説明する．次に，実験設定と結果，および結果に対する考察について述べ，最後に本研究のまとめと今後の課題について議論する．

2　関連研究

Polo ら [1] は，tinyBenchmarks と呼ばれる手法を提案し，多数モデルの応答から学習した IRT ベースの推定器を用いることで，新規モデルに対してごく少数の応答結果から能力およびベンチマーク全体スコアを高精度に推定可能であることを示した．Liao ら [2] は，正誤判定から連続スコアまで、評価の形式が違っても一つのモデルで一貫して扱える LEGO-IRT を提案しており，モデル能力を汎用能力とベンチマーク固有のオフセットに因子分解して相関構造を活用することで，少量の項目からでも安定した能力推定が可能であることを示した．これらの手法はあくまで統計的な推定器を用いたスコア予測に主眼を置いており，ベンチマークそのものの構成の最適化や，推定器を用いずに利用可能な高品質なサブセットの構築については議論の余地が残されており，本研究ではこの点に着目する．

IRT の NLP タスクへの適用については，Lalor ら [3] や Rodriguez ら [4] が，その評価における有用性を議論している．特に Castleman ら [5] は GSM8K を含む数学タスクのベンチマークに対して IRT を適用し，詳細な分析を行っている．彼らは分析の一環として識別力に基づく項目抽出も試みている一方，項目抽出の有用性等についての分析は行なっていない．本研究においてはこの点をさらに掘り下げ，項目抽出の有用性の検証を行う．

IRT によるベンチマーク構築の試みとしては，Lalor ら [6] が，確率モデルへの適合度が低い項目を除外することで，信頼性の高い静的な評価尺度を構築している．また，Li ら [7] は IRT とコンピュータ適応型テストを組み合わせたフレームワークで，各モデルに対して質の高い項目を動的に選択する方法を提案している．

3 準備

3.1 項目反応理論

項目反応理論 (IRT) は，テスト項目に対する正誤応答のパターンに基づき，被験者の持つ潜在的な能力と項目が持つ特性を共通の尺度上で統計的に推定する理論的枠組みである．

正答率に基づく古典的テストによる評価方法では，評価値である正答率が用いる項目セットの難易度に依存するという課題がある．これに対し IRT では，項目の困難度や識別力といった特性と被験者の能力を分離して推定することが可能であり，項目群の特性によらない能力の評価が実現される．

3.2 項目特性関数

IRT では，潜在特性 θ を持つ被験者の正答する確率が**項目特性曲線 (ICC)** に従うと仮定し，各種パラメータを推定する．本研究では，パラメータとして項目の困難度 β と識別力 α を持つ 2 パラメータロジスティックモデル（2PL モデル）を ICC として用いる．項目 j に対する ICC は次のように表される：

$$p_j(\theta) = \frac{1}{1 + \exp(-\alpha_j(\theta - \beta_j))}. \tag{1}$$

ここで，$p_j(\theta)$ は潜在特性 θ を持つ被験者が項目 j に正答する確率を表す．

ICC において，困難度 β_j は正答確率が 0.5 となる能力値，つまり曲線の左右位置を決定する．また，識別力 α_j は変曲点における曲線の傾きを決定する．これは，識別力が高い項目ほど特定の能力境界において被験者間の能力差を鋭敏に判別可能であることを意味する．項目パラメータの値による ICC のグラフ形状の変化を図 1 に示す．

3.3 IRT における情報量

項目情報関数 (IIF) は，ある能力値 θ に対して，各項目が能力推定の精度にどの程度寄与するかを示す指標である．また，特定の θ に対して項目情報関数が返す値は**項目情報量**と呼ばれる．2PL モデルにおける項目 j の IIF は次のように表される：

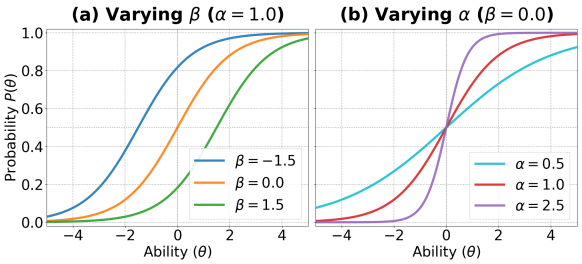

図 1: 項目特性曲線における各パラメータの影響：(a) $\alpha = 1.0$ に固定した際の困難度 β による変化 (b) $\beta = 0.0$ に固定した際の識別力 α による変化．

$$I_j(\theta) = \alpha_j^2 p_j(\theta)(1 - p_j(\theta)). \tag{2}$$

全項目の情報量の総和は，**テスト情報関数 (TIF)** と呼ばれ，特定の θ に対してテスト情報関数が返す値は**テスト情報量**と呼ばれる．2PL モデルにおける TIF は次のように表される：

$$I(\theta) = \sum_{j=1}^{n} I_j(\theta). \tag{3}$$

真の能力値と能力推定値の標準誤差 $SE(\theta)$ は情報量の逆数の平方根，$1/\sqrt{I(\theta)}$ に等しい．すなわち，$I(\theta)$ が大きいほど能力推定の精度は高まると言える．式 (2) より，テスト情報量は識別力の二乗に比例するため，識別力の高い項目を選択することがテスト情報量の最大化に直結する．一方で，式 (1) を考えると，個別の項目情報は $\theta = \beta_j$ の近傍で最大となるため，項目の困難度分布は TIF が高い値を保持する能力帯を決定する．

4 手法

本研究では，2PL モデルに基づいて推定された，ベンチマーク内各項目の困難度 β_j および識別力 α_j に基づき項目抽出を行い，抽出した項目のみによる能力推定の精度を，複数モデル間の順位付けの安定性を尺度として検証する．このプロセスは，(1) 全項目を用いたパラメータ推定，(2) 特定の指標に基づく項目抽出，(3) 抽出後のデータセットによるパラメータ再推定の 3 つのフェーズから構成される．

4.1 パラメータに基づく項目抽出

全 n 個の項目からなる集合を $\mathcal{T}_{all} = \{t_j\}_{j=1}^{n}$ とし，項目 t_j に対する困難度を β_j とする．このとき，困難度の上位 k% の項目を抽出したデータセット $\mathcal{T}_{\beta}^{k}$

を以下のように定義する．

$$\mathcal{T}_\beta^k = \{t_j \in \mathcal{T}_{all} \mid \beta_j \text{ is among the top } k\% \text{ values}\} \quad (4)$$

ここで抽出された $\mathcal{T}_\beta^k$ を用いて再度 IRT に基づく能力値の推定を行い，各モデルの潜在能力 θ_β^k を算出する．

同様に，識別力 α_j に基づく項目抽出データセット $\mathcal{T}_\alpha^k$ を次式で定義し，潜在能力 θ_α^k を算出する．

$$\mathcal{T}_\alpha^k = \{t_j \in \mathcal{T}_{all} \mid \alpha_j \text{ is among the top } k\% \text{ values}\} \quad (5)$$

このように，一定以上のパラメータ特性を有する項目のみを保持した条件下で各モデルの能力を再推定することで，項目削減が順位の安定性に与える影響を評価する．

4.2 評価指標

各抽出条件下で得られた能力推定値 θ_β^k および θ_α^k と，全項目を用いた際の能力推定値 θ との間の順位一貫性を **スピアマンの順位相関係数** ρ により評価する．各モデル i に対する 2 通りの順位をそれぞれ X_i および Y_i とし，モデル数を N 個とすると，ρ は次式で定義される：

$$\rho = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}}. \quad (6)$$

ここで，$\bar{X}, \bar{Y}$ はそれぞれの順位の平均値である．順位相関係数 ρ は -1 から 1 の範囲を取り，1 に近いほど両順位が一致していることを示す．本研究においては，任意の 2 つの（サブセットを含む）ベンチマークから得られた順位をそれぞれ X, Y に割り当てて算出を行う．これにより，項目削減による順位安定性について定量的に判定する．

なお，IRT における尺度の不確定性を考慮し，本研究では推定値の絶対誤差ではなく，実用上の観点から重要なモデル間の順位安定性に焦点を当てて評価を行う．

5 実験

5.1 設定

本研究では，分析対象のベンチマークとして数学的推論能力を測定する 8,792 個の項目からなる GSM8K [8] を用いた．項目例を付録 A に示す．実験には，Open LLM Leaderboard Archive [9] にて公開されている各項目に対するモデルの応答ログを利用した．したがって，評価対象は同アーカイブに結果が掲載されているモデルに限定される．この実験では，そのうち GSM8K に対する正答率が 3% 以上，かつ応答結果の読み取りに成功した 5,347 個のモデルを対象とした．また，Open LLM Leaderboard Archive では，GSM8K に対するプロンプトとして 5-shot の Chain-of-Thought (CoT) を用いている．

IRT のパラメータ推定には Python の pyirt ライブラリ[1)]を，スピアマンの順位相関係数の算出には SciPy ライブラリ[2)]をそれぞれ用いる．pyirt における引数設定などは，付録 B に記載する．

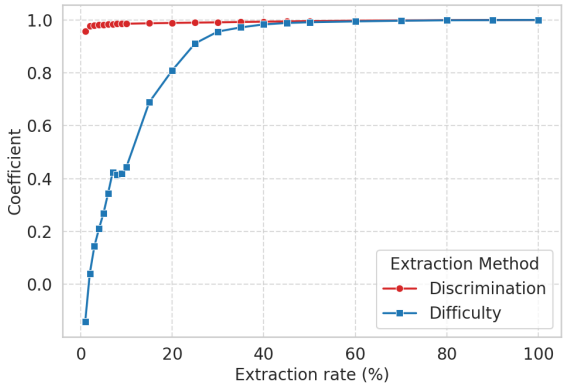

図 2: GSM8K に対するスピアマンの順位相関係数 ρ の変化．青は困難度に基づく項目抽出，赤は識別力に基づく項目抽出．横軸は抽出後のデータセットの割合．抽出は 10% までは 1% 刻み，10% から 50% までは 5% 刻み，それ以降は 10% 刻みで行なった．

5.2 結果と考察

図 2 に，GSM8K の抽出データの，元データに対するスピアマンの順位相関係数 ρ の変化を示す．識別力に基づく抽出では，1% を抽出した時点で係数が 0.95 を超え，3% 以降はほぼ等差的に増加している．一方，困難度に基づく抽出では係数がかなり低い値から開始し，識別力による抽出と同程度の相関となるまでには 40% の項目が必要となる．

図 3 に，困難度と識別力についてそれぞれ上位 10% の項目を抽出した際の潜在能力によるモデルの順位と，GSM8K 全体での順位についての散布図を示す．困難度に基づく抽出では図中の広い範囲に点が散っていることから，各モデルの能力の推定がうまくできていないことが確認できる．識別力に基づく抽出では，比較的全体での順位と近い分布となっており，順位の高低によらず安定した識別が可能で

1） https://github.com/17zuoye/pyirt
2） https://github.com/scipy/scipy

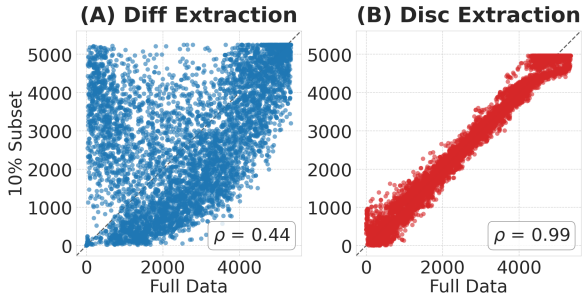

図 3: GSM8K 全体での順位と，(A) 困難度に基づく上位 10% 項目抽出，(B) 識別力に基づく上位 10% 項目抽出による順位の散布図.

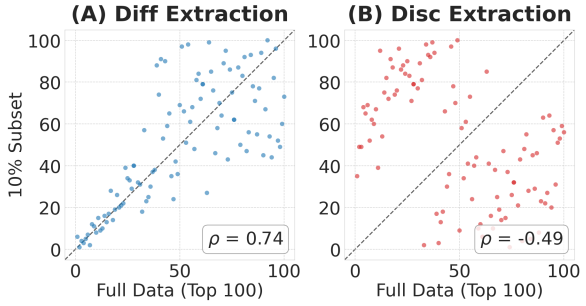

図 4: 全体での順位上位 100 モデルに絞って示した，GSM8K 全体での順位と，(A) 困難度に基づく上位 10% 項目抽出，(B) 識別力に基づく上位 10% 項目抽出による順位の散布図.

あることが示唆される.

図 4 に，全体での順位上位 100 モデルに絞った際の同様の散布図を示す．図 3 とは異なり，困難度による抽出の相関が強く，逆に識別力による抽出では負の相関となっていることが確認できる.

図 5 に，各抽出手法に基づく IRT 実行時の TIF を示す．TIF のピーク位置については 2 つの抽出手法の間で大きな差はない．一方で，潜在能力の分布については，困難度による抽出では低い能力帯に偏っていてかつ TIF のピークと乖離しているのに対し，識別力による抽出では比較的広範な能力帯にわたって分布している.

以上の結果から，困難度に基づく抽出では，高性能モデル群に対しては精度良く能力の上下を識別できる一方で，モデル群全体に対しは能力の差をうまく捉えきれないことが示唆される．原因としては，中〜低能力帯のモデルが解ける項目が少なく，まぐれでの正解の影響が大きくなり，それが能力推定時のノイズとなっている可能性が考えられる．一方，識別力に基づく抽出では，広範な能力帯で一定の情

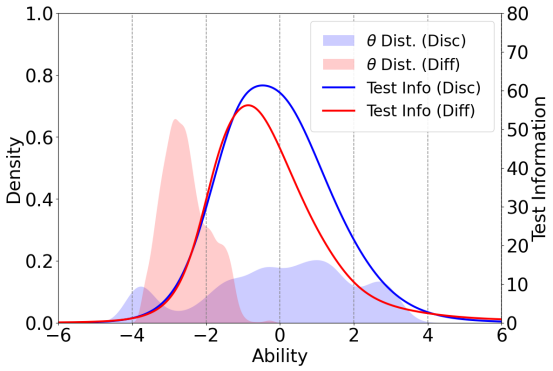

図 5: 実線：困難度・識別力に基づく上位 10% 項目抽出時の TIF，領域：同条件下での潜在能力の分布．それぞれ赤は困難度に基づく抽出，青は識別力に基づく抽出.

報量を確保できるため，元のデータセットに対して高い順位相関を維持できると考えられる．10% 程度のデータ量であっても，元データと高い順位相関を持つ能力推定が可能であり，より効率的な評価が実現の可能性が示唆される．一方，高性能モデル群に対しては性能差をほとんど捉えられておらず，局所的な能力帯に対する順位付けには不向きである可能性がある.

6 おわりに

本稿では，IRT に基づく項目特性を用いたベンチマーク項目の選択的抽出手法について GSM8K を用いて評価し，結果の分析を行った．その結果，各抽出方法について，以下のことがわかった

- 困難度に基づく抽出では，高性能モデル群に対して適した項目を提供できる一方で，モデル群全体に対しては順位相関が弱く，能力の差をうまく捉えきれない可能性がある.
- 識別力に基づく抽出では，10% 程度のデータ量であっても，大域的には元データセットと高い順位相関を持つ能力推定が可能だが，高性能モデル群に対する局所的な順位付けには不向きである可能性がある.

今後の課題としては，実験を通して得た新たな仮説の検証や，他モデル・ベンチマークへ適用することによる手法の一般性の確認が挙げられる．特に，リーダーボードの都合上，最先端モデル・ベンチマークに対する実験が行えていないため，今後の研究でこれらに対する評価を行うことが重要と考えられる.

謝辞

本研究は，JST CREST JPMJCR2565 の支援を受けたものです．

参考文献

[1] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating LLMs with fewer examples. In **Proceedings of the 41st International Conference on Machine Learning**, pp. 34303–34326, 2024.

[2] Lele Liao, Qile Zhang, Ruofan Wu, and Guanhua Fang. Toward a unified framework for data-efficient evaluation of large language models. **arXiv preprint arXiv:2510.04051**, 2025.

[3] John P. Lalor, Pedro Rodriguez, João Sedoc, and Jose Hernandez-Orallo. Item response theory for natural language processing. In **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts**, pp. 9–13, 2024.

[4] Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change how we evaluate? In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4279–4288, 2021.

[5] Jane Castleman, Nimra Nadeem, Tanvi Namjoshi, and Lydia T. Liu. Rethinking math benchmarks for llms using irt. In **Proceedings of the IRAISE Workshop at AAAI 2025**, Vol. 273 of **Proceedings of Machine Learning Research**, pp. 66–82, 2025.

[6] John P. Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 648–657, 2016.

[7] Peiyu Li, Xiuxiu Tang, Si Chen, Ying Cheng, Ronald Metoyer, Ting Hua, and Nitesh V. Chawla. Adaptive testing for LLM evaluation: A psychometric alternative to static benchmarks. **arXiv preprint arXiv:2511.04689**, 2025.

[8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. **arXiv preprint arXiv:2110.14168**, 2021.

[9] Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard. https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard, 2023.

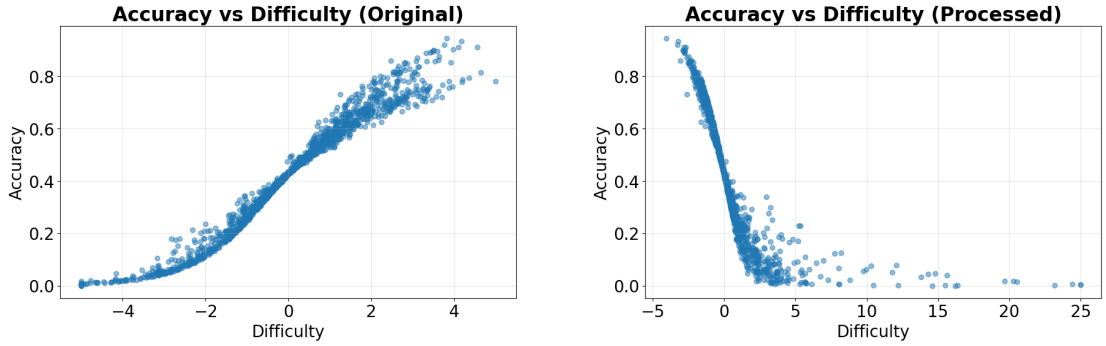

図 6: GSM8K に対する困難度と正答率の散布図．左図は処理前，右図は処理後．

A　項目例

GSM8K に含まれる項目の例を表 1 に示す．なお，記載されている困難度・識別力は GSM8K 全体に対し IRT を実行した際のもので，困難度は後述の処理を経た値である．

B　pyirt の実行設定

pyirt における IRT パラメータ推定の設定は以下の通りである．

- model_spec = ’2PL’
  使用する ICC モデルの指定
- theta_bnds = [-5, 5]
  潜在能力の上限と下限
- num_theta = 81
  パラメータ推定の際の潜在能力の近似グリッド数
- alpha_bnds = [0.2, 5]
  識別力の下限と上限　下限を 0 にすると収束しない場合があるため注意
- beta_bnds = [-5, 5]
  困難度の下限と上限
- max_iter = 1000
  推定の最大反復回数
- tol = 1e-4
  収束判定の閾値

実行の際には，可能な場合は並列処理を有効にして計算を行なった．

また，pyirt によって算出された困難度パラメータは，一般的な IRT の定義でいう $-\alpha\beta$ に相当するため，IRT でのパラメータ推定で得た各困難度パラメータについて，対応する識別力パラメータで除算を行い，符号を反転させた値を最終的な困難度パラメータとして用いた．したがって，実際の困難度パラメータが取りうる範囲は $[-5, 5]$ ではなく，$[-25, 25]$ となる．図 6 に，処理前の困難度と正答率の散布図を，図 6 に，処理後の困難度と正答率の散布図を示す．

表 1: GSM8K に含まれる項目の例

項目 1（困難度：15.62，識別力：0.2000）
Question:
Carlos is planting a lemon tree. The tree will cost $90 to plant. Each year it will grow 7 lemons, which he can sell for $1.5 each. It costs $3 a year to water and feed the tree. How many years will it take before he starts earning money on the lemon tree?

Answer:
He makes $10.5 selling lemons each year because 7 x 1.5 = «7×1.5=10.5»10.5
He earns $7.5 each year from the lemon tree because 10.5 - 3 = «10.5-3=7.5»7.5
It will take 12 years to earn enough to pay off the tree because 90 / 7.5 = «90/7.5=12»12
He will make money in year 13 because 12 + 1 = «12+1=13»13
13

項目 2 （困難度：0.85，識別力：3.33）
Question:
The Llesis family drove and hiked 6 hours to their vacation spot. They drove an average of 50 miles per hour and hiked an average of 5 miles per hour less than half their speed when they drive. If it took them 1.5 hours to hike, how far was their vacation spot?

Answer:
Half the average speed of Llesis family driving is 50/2 = «50/2=25»25 miles per hour.
Hence, their average speed hiking was 25 - 5 = «25-5=20»20 miles per hour.
So, they hiked 20 x 1.5 = «201.5=30»30 miles.
It took them a total of 6 - 1.5 = «6-1.5=4.5»4.5 hours driving.
So, they drove 50 x 4.5 = «504.5=225»225 miles.
Therefore, their vacation spot was 30 + 225 = «30+225=255»255 miles far.
255