

目標準拠テストの作成事例

田中 利一郎* 山際 一朗* 生田 孝至*
高橋 一久** 丸山 裕輔** 横尾 浩** 沢栗 美香子**

<概 要>

小学校第3学年の算数科の単元「分数」の一部について、児童の到達度をみるためのテストを作成した。これを79人の児童に実施し、S-P表分析、及び、同じ目標を反映することを意図して作成した項目どうしの結果の一致度をもとにして、テスト中に不適当な項目がないかをみた。不適当と思われる項目を除いた上で、任意に設定した分割点による決定の一致度をもとに、信頼性を推定した。

I はじめに

目標に準拠した評価（測定）は、1960年代後半から議論されてきた。これは、明確に定義された学習課題について、児童・生徒の到達の度合いを記述するのに有効な評価法である。日本でも、近年「到達度評価」の名は一種の流行のごとく広く知られるようになったが、技術的な研究は積まれているのが現状であるといえるだろう。

目標に準拠した評価は、(1)各目標領域の具体化表の作成、(2)各目標領域をうまく代表する項目サンプルの選出、(3)そのサンプルによる児童・生徒の得点に基づいて到達度を判定するための分割点の設定という手続きを踏む（橋本，1981）。この種の評価に関する従来の実践報告は、(1)の目標の具体化のみを扱うものが多いようである。そこで本研究は、(2)の項目のサンプルの選出も扱い、信頼性の推定といった問題も考察することを意図して行った。

II 目 的

本研究では、小学校算数科の一単元（部分）を取り上げ、次のような目的を設定した。

- (1) 取り上げた単元（部分）において、児童の到達度を測定するための目標準拠テストを作成すること。
- (2) テスト結果から、作成したテストの項目と信頼性を検討すること。

* 新潟大学教育学部

** 新潟大学大学院教育学研究科

Ⅲ 手 順

本研究の枠組みを、図に表すと以下のようなになる。

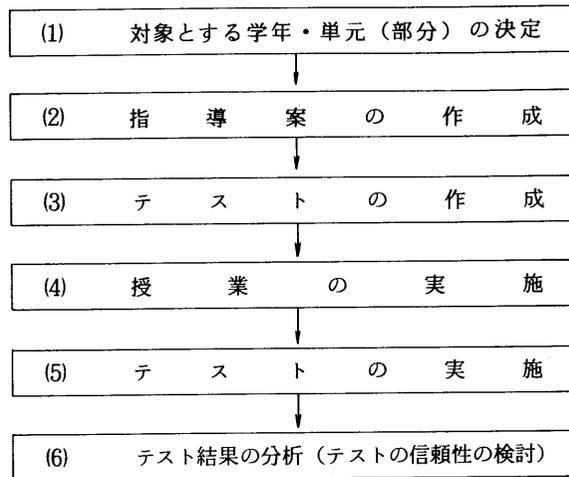


図 1 本研究の枠組み

この枠組みに沿って、本研究の手順を具体的に述べる。

最初に、対象とする学年と単元を、次のように決定した。対象とする学年は、本研究に協力して下さる先生方の担当学年である第3学年に決定した。対象とする単元は、第3学年で児童がはじめて学習する「分数」の単元に決定した。本研究では、その単元の中で導入にあたる分数の概念を指導する部分を取り上げた。

2番目に、授業を実施するための指導案の作成を行った。指導案作成の最初の段階は、1時間目から5時間目までの各時の目標（教授目標）の設定である。各時の目標は、次のような手順を経て設定した。まず、取り上げる単元の部分に関して、教科書の比較を行った。つまり、各教科書に記述されている学習内容を1項目の学習内容になるようにして、それを目標行動の形でカードに記述した。その後、カードの中で、似通った学習内容が記述してあるカードを、ひとまとまり（セット）にした。次に、カードのまとまり（セット）を単元全体の指導計画に沿って配置し、指導内容のアウトラインを決定した。そのアウトラインに基づいて、目標行動の形式で教授目標を設定した。5つの教授目標を設定したので、分数の概念を指導する部分に対して授業の配当時間を5時間とした。それに続いて第6時間目を目標準拠テスト実施の時間とした。以下に、設定した各時の目標（教授目標）を示す。

表 1 各時の目標（教授目標）

時 間	目 標
1時間目	$\frac{1}{2}$ 、 $\frac{1}{3}$ 、 $\frac{1}{4}$ の意味と表し方がわかる。
2時間目	既習の単位に満たない量を、その単位と単位分数を用いて表すことができる。
3時間目	1 mをこえる長さを、1 mと $\frac{1}{10}$ mと表すことができる。
4時間目	単位分数を集めた大きさを、分数で表すことができる。
5時間目	「分数」「分母」「分子」の用語とその意味がわかる。

各時の目標を設定した後、1時間目から5時間目までの指導案を作成した。指導案は、協力校の先生方とともに、話し合いで検討し、修正した。

3番目に、児童の到達度を測定するための目標準拠テストの作成及び検討を行った。まず、作成した指導案をもとにして、最初の5時間で指導する目標領域を書き出した。そして、その目標領域に対応するテスト項目を作成した。さらに、そのテスト項目が、問おうとすることを妥当に反映できそうかどうかを、協力者の先生方とともに、話し合いで検討した。

4番目に、新潟市内のA小学校・B小学校（各1クラス）において、クラス担任の先生（本研究協力者）により、授業（1時間目～5時間目）を実施した。授業は、事前に作成した同一の指導案を用いて、実施された。

5番目に、作成した目標準拠テストを、各クラスにおいて、6時間目に実施した。

なお、授業とテストは、昭和60年7月上旬から中旬にかけて実施された。

最後に、テスト結果の分析を行った。ここでは、実施されたテストに対して、S-P表による分析と項目反応の一致度から、不相当だと思われる項目を除いた上で、その信頼性を推定した。

続いて、本研究の中心であるテストの作成とテスト結果の分析について述べる。

IV テストの作成

児童の到達度を測定するための目標準拠テストは、次に述べるように、2段階を経て作成した。

最初の段階は、目標領域の設定である。テストの目標領域を、作成した5時間分の指導案をもとにして、以下のように設定した。各目標領域は、指導の過程（授業の流れ）にほぼ沿ってあげた。

- A. ある大きさを○等分した大きさを、 $\frac{1}{10}$ とすることができる。
- B. ある大きさを等しく○つに分けていないと、 $\frac{1}{10}$ とはいえないことがわかる。
- C. ある大きさを○等分したうちの△つ分の大きさを、 $\frac{\triangle}{10}$ とすることができる。
- D. もとの大きさが異なると、同じ分数の大きさが異なるということがわかる。
- E. 求める大きさの位置が変わっても、ある大きさを○等分したうちの△つ分の大きさを $\frac{\triangle}{10}$ とすることができる。
- F. 同じ大きさでも、もとの大きさが異なれば、その大きさを表す分数も異なるということがわかる。
- G. 長さなどの単位の決まった大きさを表すには、分数とその単位を用いて表すということがわかる。

- H. 数直線の1までの長さをもとの大きさとして、数直線上の分数をいうことができる。
- I. 数直線を用いて、ある単位をもった大きさを、分数でいうことができる。
- J. 単位量を越えた大きさを、整数と分数でいうことができる。
- K. $\frac{\triangle}{\bigcirc}$ は、 $\frac{1}{\bigcirc}$ を \triangle つ集めた分数だということができる。
- L. $\frac{\bigcirc}{\bigcirc}$ は、1と同じ大きさであるということができる。
- M. $\frac{\triangle}{\bigcirc}$ は、 $\frac{1}{\bigcirc}$ の \triangle 倍の大きさであるということができる。
- N. $\frac{\triangle}{\bigcirc}$ のように表した数を、「分数」ということができる。
- O. $\frac{\triangle}{\bigcirc}$ の \bigcirc にあたる数を、「分母」ということができる。
- P. $\frac{\triangle}{\bigcirc}$ の \triangle にあたる数を、「分子」ということができる。
- Q. 分母は、1を何等分したかを表していることがわかる。
- R. 分子は、単位分数がいくつ分あるかを表しているということがわかる。

第2段階は、目標領域に対応するテスト項目の作成である。まず、授業で取り上げた問題の形式に沿う形で（授業での発問のいまわしとテスト項目の問かけが一致するように考慮して）、テスト項目を作成した。その際、テスト項目が目標領域の内容を満遍なく反映するように配慮した。次に、作成したテスト項目を、協力者の先生方とともに、話し合いで検討した。話し合いによって欠点があると思われたテスト項目を修正し、テスト項目の配列の仕方などを考慮して、テストを構成した。なお、後にテストを検討する際、折半法を用いることをあらかじめ考えていたので、同じ内容を問うと考えられるテスト項目を、2問ずつ設定した。

このようにして作成したテスト問題を、資料1に示す。また、そのテスト項目と目標領域の対応を表2に示す。

表2 テスト項目と目標領域との対応

項目	目標領域	項目	目標領域	項目	目標領域
1-1	A, B	3-3	C, G	5-5	L
1-2	A, B	3-4	C, G	5-6	M
2-1	A, D	4-ア	A, H	6-1	A, H, I
2-2	A, D	4-イ	C, H	6-2	C, H, I
2-3	C, F	4-ウ	A, H	7-1	C, H, I
2-4	C, F	4-エ	C, H	7-2	C, H, I, J
2-5	C, E	5-1	K,	8-1	N
2-6	C, E	5-2	L,	8-2	O
3-1	A, G	5-3	M,	8-3	P
3-2	A, G	5-4	N,	8-4	Q
				8-5	R

V テスト項目の検討

1 S-P表による分析

S-P表は個々の児童の学習診断に用いられることが多いようだが、ここでは、不適切な項目を見つけ出すためにS-P表を用いた。特に正答率と注意係数をもとにして検討すべき項目を特定しようとした。

注意係数*とは、S-P表において、その変量の区切線（項目についてはP曲線）を境に1（正答）と0（誤答）の入れ変わりのない完全反応パターンからみて、実際に得られた反応パターンがどの程度異なっているかを数量化したものである。数量化においては、実際の反応パターンと完全反応パターンの差異を、完全反応パターンと無作為反応パターンの差異で基準化している(佐藤,1975,1985a)。

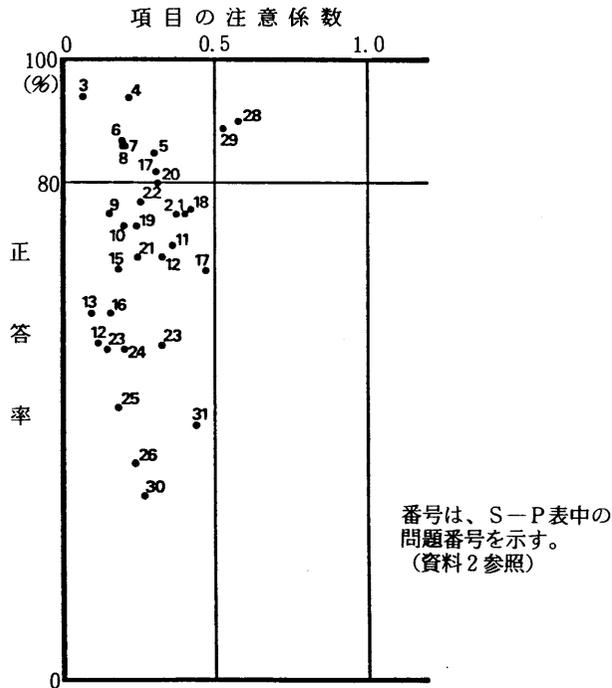


図1 各項目の注意係数と正答率

$$* \text{ 注意係数} = \frac{\left(\begin{array}{l} \text{完全な反応パターンと} \\ \text{基準変量との共分散} \end{array} \right) - \left(\begin{array}{l} \text{実際の反応パターンと} \\ \text{基準変量との共分散} \end{array} \right)}{\left(\begin{array}{l} \text{完全な反応パターンと} \\ \text{基準変量との共分散} \end{array} \right) - \left(\begin{array}{l} \text{無作為な反応パターンと} \\ \text{基準変量との共分散} \end{array} \right)}$$

項目の注意係数の場合、基準変量は生徒の合計得点である。

この式を変形すると、項目の注意係数は、

$$\frac{\left(\begin{array}{l} \text{その項目のP曲線から上の“0”に} \\ \text{対応する合計得点の和} \end{array} \right) - \left(\begin{array}{l} \text{その項目のP曲線から下の“1”に} \\ \text{対応する合計得点の和} \end{array} \right)}{\left(\begin{array}{l} \text{その項目のP曲線から上の} \\ \text{生徒の合計得点の和} \end{array} \right) - \left(\begin{array}{l} \text{その項目の正答者数} \\ \times \text{ (平均得点)} \end{array} \right)}$$

で得られる。

つまり、注意係数の高い項目は、S-P表全体の反応パターンを基準としたとき、異質な項目であることになる。したがって、注意係数を不適切な要素を含む可能性のある項目を見出す手がかりとして用いることができる。

作成したテストを実施し、解答を正答か誤答かの2値データに直した上で、その結果からS-P表を作成した。これを資料2に示す。また、図1は、正答率と注意係数に基づいて、各項目をプロットしたものである。これを見ると項目7-1(25)、7-2(26)、8-4(30)、8-5(31)の正答率が低く、8-2(28)と8-3(29)の注意係数が比較的高いことがわかる。(かっこ内は図1における番号を示す)。

2 項目の一致度による分析

同質の項目に対して、同じ児童が異なる反応を示した場合、原因の1つとして、項目そのものが不適切であったことが考えられる。したがって、項目間の反応の一致をみることで、不適切な要素を含む項目を見出す手がかりを得ることができる。

ここでは、実施したテストのうち、項目8-1を除く30項目を15項目ずつに折半して2つのテスト(テストI、テストIIと呼ぶ)とし、それらの項目間で反応が一致している程度をみることにする。折半は、テスト作成時に同じ内容を問うものとして設定した2つの項目を、テストIとテストIIにランダムに振り分けて行った。

振り分けた項目のペアについて、同じ児童が同じ反応を示しているかどうかを調べた。各反応の度数と、これに基づく各項目の反応の一致度を表3に示す。これを見ると、多くの項目のペアが、0.8以上の一致度を示すが、項目8-5と8-4のペアの一致度が0.696と際立って低い。また、5-5と5-2のペアも0.747と低い値を示している。

表3 折半した項目とその一致度

項目		度数				項目の一致度
テストI	テストII	F(0, 0)	F(0, 1)	F(1, 0)	F(1, 1)	
1-1	1-2	23	1	1	54	0.975
2-1	2-2	9	1	1	68	0.975
2-4	2-3	15	0	2	62	0.975
2-6	2-5	16	0	0	63	1
3-2	3-1	20	2	1	56	0.962
3-3	3-4	21	3	4	51	0.911
4-U	4-A	26	1	6	46	0.911
4-I	4-E	29	7	2	41	0.886
5-4	5-1	9	7	5	58	0.848
5-5	5-2	12	13	7	47	0.747
5-6	5-3	14	9	6	50	0.810
6-1	6-2	33	4	4	38	0.899
7-1	7-2	42	2	9	26	0.861
8-2	8-3	7	1	2	69	0.962
8-5	8-4	39	8	16	16	0.696
全	体	315	59	66	745	0.895

$F(t_1, t_2)$:
 テストIで t_1 、テストIIで t_2
 の反応をした度数
 t_1 : テストIの反応
 t_2 : テストIIの反応
 t_1 と t_2 は、1か0をとる。
 1は正答、0は誤答を表す。

項目の一致度：

$$\frac{F(0, 0) + F(1, 1)}{\sum_{t_1} \sum_{t_2} F(t_1, t_2)}$$

3 項目の検討

S-P表による分析と項目の一致度による分析で、正答率や一致度が低かった項目や、注意係数が高かった項目について、不適切な点がないか話し合いで検討した。

正答率の低い7-1と7-2については、この項目で問われている内容自体が難しく、誤答の状態からみて、この内容の学習が十分でなかったためと解釈した。

注意係数の比較的高かった8-2と8-3、及び一致度の低かった5-2と5-5については、児童の解答を見る限り、項目の意味が児童に伝わらなかったことはないと判断した。

正答率と一致度がともに低い8-4と8-5については、テスト作成時に予想していなかった解答が寄せられているため、作成者の意図するところが児童にうまく伝わらなかった危険が高いと判断し、この2項目は削除することにした。これらの項目を作成した時には、1番目の空欄には、「分母」「分子」という言葉か、分母や分子を示す数字が入られることを期待していた。ところが、この空欄に「数」「分数」と解答している児童がみられた。このような結果になったのは、1項目中に2つの空欄を設定するという問題作成上の基本的ミスによる。このミスによって、空欄にあてはまるべき内容が限定されなくなり、多義にとらえられやすい性質を帯びたと思われる。加えて、一文中に空欄が2つあることから、文を構成する要素間の関係がわかりにくくなったことも考えられる。

また、項目8-4については別の欠点もある。この項目は、1番目の空欄に「分母」または「3」、2番目の空欄に「3」が入られるべきものとして設定している。すると、「 $\frac{2}{3}$ の分母(3)は、1を3等分したことをあらわします」という文になる。しかし、この文は、分母についてだけ限定して考えればよいが、「1を3等分したこと」と $\frac{2}{3}$ を同等に考えようとする、食い違いが生じてしまう。8-5に正答しながら、8-4で誤答している児童が多いのは、このことによるものと思われる。

したがって、項目8-4と8-5は、児童の力を正当に評価できたとはいえないので、削除せざるを得ない。

また、項目2-1と2-2はIVであげた目標領域のA(ある大きさを○等分した大きさを $\frac{1}{\bigcirc}$ ということができる)とD(もとの大きさが異なると、同じ分数の大きさが異なるということがわかる)に対応させて作成したが、Dについては、これらの項目ではうまく反映できないようである。つまり、Aを習得しDを習得していない児童でも正答できる可能性が非常に高い。これらの項目の一方に正答し、一方に誤答している児童は2人しかおらず、いずれも一方を $\frac{1}{2}$ 、一方を $\frac{1}{3}$ と答えていた。これと同様に、項目2-3と2-4は目標領域のC(ある大きさを○等分したうちの△つ分の大きさを $\frac{\triangle}{\bigcirc}$ ということができる)とF(同じ大きさでも、もとの大きさが異なれば、その大きさを表す分数も異なるということがわかる)に対応させて作成したがFについては、うまく反映していない危険がある。これらの項目に同じ分数で答えた児童は1人であった。

したがって、項目2-1と2-2は目標Aのみを、また項目2-3と2-4は目標Cのみを反映したと解釈するのがよいと考えられる。

VI 信頼性の推定

目標準拠テストの信頼性は、同一の目標領域からサンプリングして作成した2つのテストの結果の一貫性に基づいて考えることができる。

実施したテストを折半して構成したテストIとテストIIから、それぞれ項目8-5と8-4を除く

と、2つのテストの正答率の度数分布は表4のようになる。この度数をもとに、2つのテストと正答率の分布が独立かどうかをカイ2乗検定すると、独立であると判断できる。

表4 テスト×正答率のクロス表

正答率 テスト	.0 ~	.4 ~	.6 ~	.8 ~	.9 ~	計
テスト I	12	11	13	21	22	79
テスト II	10	16	13	17	23	79
計	22	27	26	38	45	158

$$x^2 = 1.511$$

$$(df = 4, n. s.)$$

ここでは、分割点を任意に1つに設定し、それによる決定*が、テストIとテストIIの2つのテストで一致しているかどうかを各児童について調べる。このように決定の一致度を用いた方法で、かつ偶然的要因に伴う一致を除いて一致度を指数化したものとして、Swaminathan et al. (1974)の信頼性係数 \hat{k} がある。この信頼性係数を用いて、信頼性をみていくことにする。

任意の分割点を正答率にして0.8に設定したときの、2つのテストの決定の状態(4通りある)の比率を表5に示した。このとき、 \hat{k} は0.873になる。分割点を変化させると、当然ながら \hat{k} の値も変動する。0.8という分割点は任意のものなので、分割点が変わると、 \hat{k} の値がどのくらいになるかをみた。その結果を表6に示す。いずれの場合も、 \hat{k} は比較的高い値である。

表5 2つのテストにおける決定

テストII テストI	pass	fail	計
pass	0.443	0.051	0.494
fail	0.013	0.494	0.506
計	0.456	0.544	1

$$\hat{k} = \frac{P_o - P_c}{1 - P_c}$$

$$P_o = \sum_i P_{ii}$$

$$P_c = \sum_i P_i \cdot P_i$$

ただし
 P_{ii} : テストIで状態i、
 テストIIで状態iである比率。

$$\hat{k} = 0.873$$

表6 分割点による \hat{k} の変化

分割点	0.5	0.6	0.7	0.8	0.9
\hat{k}	0.835	0.913	0.865	0.873	0.753

Ⅶ 結 語

到達度テストを用いて、児童・生徒の学習診断を行っている実践報告は数多く見受けられるが、それらがテストの作成過程やテスト項目の良否を考慮して行われているかどうかは疑問に思われる。テスト結果に基づく判断を確かなものにするためにも、教育実践において、テストの作成過程やテスト項目に一層の配慮がなされるべきであろう。

* 分割点が1つするとき、起こりうる決定は、分割点に到達した場合と、到達しなかった場合の2通りである。ここでは、前者をpass、後者をfailと呼ぶ。

私達は、目標領域を確定した上で、それを反映する項目を作成し、このようにして作成した項目を配列してテストを作成するという手続きを踏むことによって、テストの内容的妥当性を確保しようとした。目標領域の設定や項目の作成では、現場（小学校）の先生方も交えて話し合いを積んだつもりであったが、実際のテスト結果によって、項目に不適当な点のあったことが指摘されたわけである。テスト作成時に、目標領域がうまく設定できているかどうかや、作成した項目が目標をうまく反映できそうかどうかを、教科の専門家やテストの専門家に判断してもらえば、さらにテストを改善できたであろう。

テストの結果から、個々の児童・生徒について到達度を判断する際には、到達したかどうかを決定する分割点をいかにして設定するかということが重要な問題となる。分割点は、従来経験的に定められてきたが、近年は、海外の研究者によって分割点を組織的な手続きを用いて定める手法がいくつも提案されてきているようである。しかし、本研究では、分割点の設定の問題まで触れることはできなかった。

なお、本稿は、新潟大学大学院教育学研究科の昭和60年度の演習「授業の設計と分析」で取り上げた課題について、新潟教育学会第25回大会で発表した内容をまとめたものである。

文 献

- 1) Carmines, E. G. & Zeller, R. A. 1979 *Reliability and validity assessment*. SAGE Publications. (水野欽司・野嶋栄一郎訳 1983 テストの信頼性と妥当性 朝倉書店)
- 2) 橋本重治 1981 到達度評価の研究 図書文化
- 3) Millman, J. & Popham, W. J. 1974 The issue of item and test variance for criterion-referenced test: A clarification. *J. educ. Meas.*, 11, 137-138.
- 4) 野嶋栄一郎 1979 到達基準による測定・評価と結果処理 梶田叡一編著 到達度評価の理論と教育革新 明治図書, 98-117.
- 5) 佐藤隆博 1975 S-P表の作成と解釈 明治図書
- 6) —— 1982 S-P表の見方と学習診断のための利用法 佐藤隆博編著 S-P表の活用 (小学校編) 明治図書, 8-29.
- 7) —— 1984a 米国におけるS-P表分析の研究・利用の現状 指導と評価, 30(9), 44-47.
- 8) —— 1984b 形成的評価におけるS-P表分析の利用 指導と評価, 30(10), 38-41.
- 9) —— 1984c 誤答内容のタイプを表示したS-P表 指導と評価, 30(11), 38-41.
- 10) —— 1984d 観点別学習診断とカテゴリー化S-P表 指導と評価, 30(12), 37-40.
- 11) —— 1985a 注意係数 — 項目反応パターンの分析 指導と評価, 31(1), 40-43.
- 12) —— 1985b 差異係数 — S-P表を利用するときの留意事項 指導と評価, 31(2), 39-43.
- 13) Swaminathan, H., Hambleton, R. K. & Algina, J. 1974 Reliability of criterion-referenced tests: A decision theoretic formulation. *J. educ. Meas.*, 11, 263-267.
- 14) Wood, D. A. 1960 *Test construction: Development and interpretation of achievement tests*. Charles E. Merrill. (池田央訳 1970 試験問題の作り方 日本文化科学社)

<資料 1> 実施したテスト問題

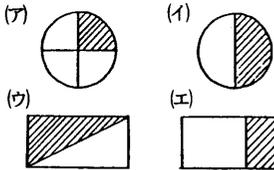


まとめのれんしゅう

くみ	ばん
なまえ	

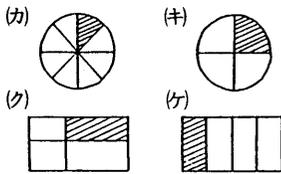
① 下のもんだいに答えましょう。

① つぎの4つの図のうち、ぜんたいの $\frac{1}{2}$ だけぬってあるものはどれでしょう。あるだけ書きなさい。



()

② つぎの4つの図のうち、ぜんたいの $\frac{1}{4}$ だけぬってあるものはどれでしょう。あるだけ書きなさい。



()

② のところは、ぜんたいのなん分のいくつですか。

① ()

② ()

③ ()

④ ()

⑤ ()

⑥ ()

③ 下のテープの長さは1mです。 のところの長さを書きましょう。

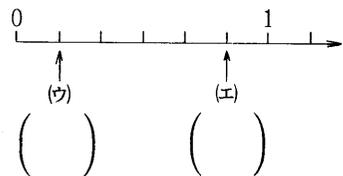
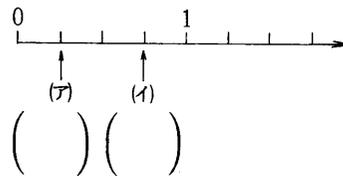
① ()

② ()

③ ()

④ ()

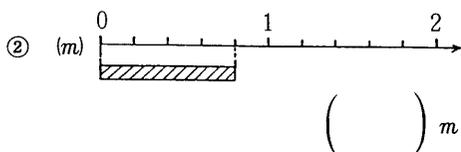
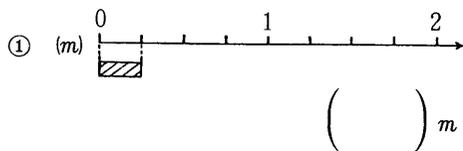
④ 下の数直線で、(ア)～(エ)はそれぞれどんな数をあらわしていますか。



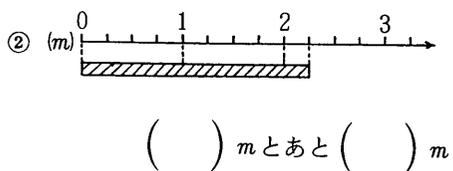
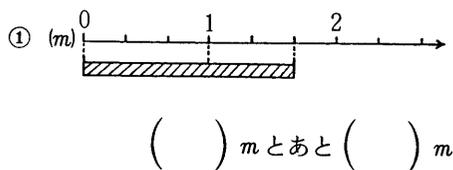
⑤ にあてはまる数をいれましょう。

- ① $\frac{7}{3}$ は、 $\frac{1}{3}$ を つあつめた数です。
- ② $\frac{1}{3}$ を つあつめた大きさは1です。
- ③ $\frac{1}{3}m$ の2ばいは、 m です。
- ④ $\frac{1}{8}$ を5つあつめた数は、 です。
- ⑤ $\frac{\text{□}}{5}$ は、1と同じ大きさです。
- ⑥ $\frac{5}{7}m$ は、 $\frac{1}{7}m$ の ばい입니다。

⑥ 下のテープの長さは、何 m でしょう。



⑦ 下のテープの長さは、何 m とあと何 m でしょう。



⑧ にあてはまることばや数をいれましょう。

- ① $\frac{1}{4}$ 、 $\frac{2}{3}$ のようにあらわした数を といいます。
- ② $\frac{1}{4}$ の分母は、 です。
- ③ $\frac{1}{4}$ の分子は、 です。
- ④ $\frac{2}{3}$ の は、1 を 等分したことをあらわします。
- ⑤ $\frac{5}{6}$ の は、 $\frac{1}{6}$ の ぶんであることあらわします。

