JSiSE Research Report
vol.31,no.5(2017-1)

テスト理論に基づいた項目分析支援システムの作問現場での活用と評価

林貴史[*1], 高木正則[*1], 山田敬三[*1], 佐々木淳[*1]
*1 岩手県立大学大学院ソフトウェア情報学研究科

Effectiveness Evaluation of a Test Item Analysis Support System Developed by Using Test Theory

Takafumi Hayashi[*1], Masanori Takagi[*1], Keizo Ymada[*1], Jun Sasaki[*1]
*1 Graduate School of Software and Information Science, Iwate Prefectural University

In many education fields, various tests have been performed to measure the intelligence and knowledge of learners. It is difficult for test creators to evaluate that each test item was whether good or bad theoretically. In this study, in order to improve the test quality, we have developed a test item analysis support system based on test theory. The system analyzes test items based on answer data of examinees and provides appropriate advices for next test creation. In this paper, we describe the main design concepts of the developed system and its effectiveness evaluation results in a creating environment of test items for a regional knowledge certification test held in Morioka city, Iwate Prefecture.

キーワード: テスト理論，項目分析，設問回答率分析図

1. はじめに

多くの教育現場では，学習者の能力や学習効果を測定するためにテストを実施し，テストの得点から各学習者を評価する．しかし，テストに出題された問題（以下，項目）の良し悪しを評価することは少ない．その要因として，項目の分析に必要となるテスト理論や統計学などの専門知識が不足していることが考えられる．また，項目の分析結果から，次回作問時の改善点やテスト項目の難易度のバランス調整などを客観的に行うのは難しい．そこで，我々は作問者がテスト受検者の解答結果に基づいて次回作問時の改善点を把握することを目的とし，テスト理論に基づいた項目の分析結果や次回作問時のアドバイスを提示する項目分析支援システムを提案・開発した．これにより，作問経験の少ないテスト作成者でも項目を評価でき，次回作問時に作成される問題の質向上，試験全体の信頼性向上や作問者の作問スキルの向上が期待できる． 本稿では，テスト受検者の解答データに基づく項目分析結果の提示方法や，作問アドバイスの生成ルールならびに提示方法について述べる．また，岩手県盛岡市で開催されているご当地検定「盛岡もの識り検定」(以下, もりけん) の解答結果を活用し，作問現場で本システムを利用してもらった結果から，分析結果の提示が検定試験の作問をする上で有効なのかを評価する．

2. 関連研究

樋口 [(1)] は，テスト理論の知見を有さない教授者が容易にテスト理論を用いて各小問別，各受検者別のスコアデータを解析できる Web アプリケーションを開発している．基本的な統計量に加え，テスト全体についての情報である信頼性係数やテストの合計点に想定される誤差の情報を計算できる．また，熊谷 [(2)] は，項目反応理論によるテスト分析を行うソフトウェアの Easy Estimation を開発している．Easy Estimation

は研究目的に限り無料で利用できる国産のフリーソフトウェアであり，GUI（Graphical User Interface）によりマウス操作のみで分析できるため，テスト分析の入門者においても容易に分析できる．多母集団分析や一部の項目母数固定による分析など，実用上必要な分析オプションも用意されている．これらのシステムでは，テスト結果の分析といった点で本研究と類似するが，作問現場に密着した次回作問時の改善点までを対象としていない点で異なる．

3. 研究課題と課題解決へのアプローチ

本研究では，①テスト理論の知見を有さないテスト作成者でも理解できるようなテスト分析結果の表示方法の解明と，②次回テスト作成時の参考になる作問アドバイスの生成ルールの構築が研究課題となる．課題①については，項目の良し悪しを視覚的に判断できるように，良い問題と悪い問題を一覧で表示する．また，設問解答率分析図[3]（4.2 節で後述）やヒストグラムなどのグラフを活用した表示方法を検討する．課題②では古典的テスト理論によって算出される各項目のパラメータ（項目難易度，項目識別度，S-P 表分析から得られる注意係数の値など）から次回作問時における改善点（アドバイス）を生成するルールを構築する．

4. 項目分析支援システム

4.1 システムの概要

提案する項目分析支援システムの概要を図 1 に示す．項目の良し悪しは出題意図によって判断基準が異なるため，システム利用者はテスト受検者の解答データだけでなく各項目の出題意図も本システムに入力する（図 1 ①）．項目分析モジュールでは古典的テスト理論などを駆使して，各項目を分析する（図 1 ②）．分析の際には，出題意図を考慮して適切な分析方法や評価基準を採用する．例えば，授業で教えた内容のうち最も基本的で全員が理解していることを確かめる問題であれば，全員が正解しても不適切な難易度の問題とは判断しないようにする．項目説明 DB は，表 1 に示す基準（参考文献 4～8 を参考に項目の評価指標・評価手法・評価基準を設定）を元に信頼性，難易度，識別度，注意係数の各数値に対応させた項目に関する説明が登録されている．項目分析モジュールでは，解答データを分析した結果と項目説明 DB で参照した各数値の補足説明を利用者にフィードバックする（図 1 ③）．また，この分析結果と出題意図は作問アドバイス生成モジュールに渡される（図 1 ④）．作問アドバイス生成モジュールでは，作問アドバイス生成ルール DB を参照（図 1 ⑤）して生成した作問アドバイスを利用者に提示する（図 1 ⑥）．

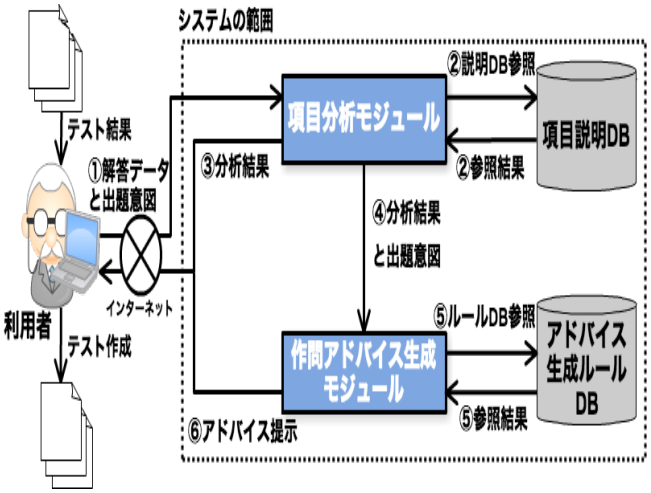


図 1 システム概要図

4.2 項目分析モジュール

（1）テスト全体の評価

一般に，テストは信頼性と妥当性によって評価されるが，妥当性は定量的評価が難しいことから，本研究ではまず信頼性を評価する機能について検討した．信頼性を評価する方法には内部一貫法，折半法，再テスト法，平行テスト法など数多くの手法がある[4]．今回はテスト全体の信頼性を評価するために一般的に広く用いられている内部一貫法のクロンバックのα係数を利用して信頼性を評価することとした．その他に，テスト全体のデータの分布状況を視覚的に認識するためのグラフとしてヒストグラムを用いた．ヒストグラムの作成には，階級数を決める必要があるため，スタージェスの公式を利用した．スタージェスの公式で階級数を決め，ヒストグラムを作成し，視覚的に受検者がどの得点にばらついているのかが分かる．

（2）各項目の評価

個々の項目の評価では，現状，利用者が入力した出題意図や各項目の予想正答率に基づく項目分析方法が未実装であるため，本稿では出題意図を考慮しない分析方法について述べる．個々の項目を評価するパラメ

ータには，古典的テスト理論で算出した難易度と識別度，S-P 表分析を元に算出した項目注意係数を利用する．現状の分析では，単一のテストを想定しており，複数のテストの分析結果を比較するような機能に関しては検討段階であるため，項目反応理論よりも分析時間が短時間で行える古典的テスト理論を用いている．分析結果は，項目説明 DB を参照し，能力を判定する上で適切と思われる問題グループ（以下，良問グループ）と不適切と思われる問題グループ（以下，悪問グループ）に分けて具体的な数値や説明とともに提示する．本システムでは表 1 の基準値に基づき難易度が 0.4 以上 0.8 未満で識別度 0.4 以上かつ注意係数 0.5 未満の項目を良問グループとした．また，難易度が 0.4 未満または 0.8 以上で識別度 0.3 未満かつ注意係数 0.5 以上の項目を悪問グループとした．良問にも悪問にも該当しない項目に関しては，標準的な問題群としてアドバイスが省略される．項目分析結果画面の最後には，良問グループ・悪問グループを含んだ全項目の設問解答率分析図と難易度･識別度･注意係数が表示される．設問解答率分析図を利用することでテストを構成している項目の良し悪しを視覚的に確認できる．特性図の作図方法は，まず受検者のテストの合計得点を昇順に並べ替え，受検者を 5 群に等分割する．5 群において最も受検者のテスト合計が高い群をレベル 5（以下，Lv5）とし，最も受検者のテスト合計が低い群をレベル 1（以下，Lv1）とする．割り切れない場合には値を繰り上げて， Lv5 から降順に割り振っていく．次に各群の正答率を計算する．縦軸に正答率，横軸に 5 群をとるグラフをプロットする．各プロットを直線で結ぶことにより作図する．作図方法に基づき，実際のデータから作図した設問解答率分析図を図 2 に示す．図 2 に示した項目は左側の解答者群（テスト全体の得点が低い解答者群）の正答率が低く，右側の解答者群（テスト全体の得点が高い解答者群）になるにつれ正答率が高くなっているため識別度が高く，測りたい特性がよく識別されている項目であることを示している．

表 1 項目の評価指標・評価手法・評価基準

指標	手法	基準
信頼性	平行テスト法 折半法	ば，信頼性が高い [5]
	内部一貫法（クロンバックのα）	“信頼性 0.71 以上”であれば，テスト全体の信頼性が高い [6] “信頼性 0.71 未満”であれば，項目数や受検者人数が少ないかテスト全体の信頼性が低い
難易度	古典的テスト理論	“難易度 0.4 未満”または”難易度 0.80 以上”の項目は難易度が不適切な項目，”難易度 0.3 未満” または”難易度 0.90 以上“は修正及び検討が必要な項目 [4]
識別度	古典的テスト理論	“識別度 0.3 未満”は識別度が低く不十分な項目，”識別度 0.2 未満”は識別度が極端に低く合否判定には直結しない項目 [4]
注意係数	S-P 表分析	“注意係数 0.5 以上 0.75 未満”は注意すべき項目 “注意係数 0.75 以上”は特に注意が必要な項目 [7] [8]

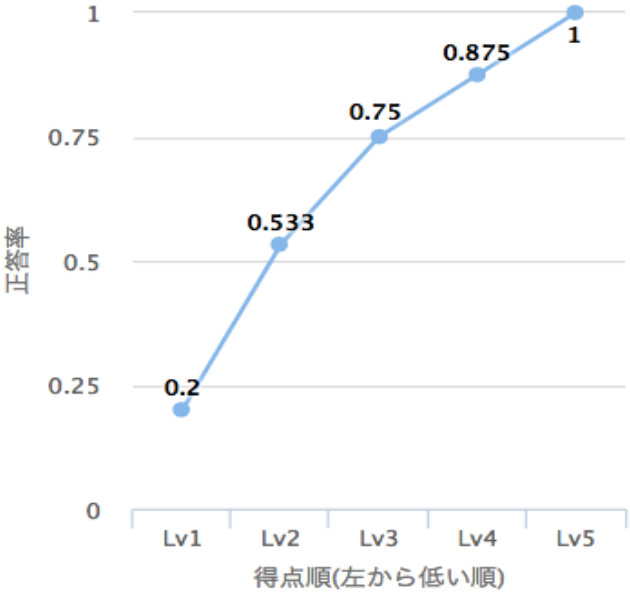


図 2 設問回答率分析図

4.3 作問アドバイス生成モジュール

図 3 に作問アドバイス生成の流れを示す. 例として, 項目番号 1 の難易度 0.26, 識別度 0.16, 注意係数 0.76 の問題がある場合, ①表 1 で示した基準値に当てはめ, 難易度/識別度/注意係数の特徴の判断を行う. 次に, ②その問題に関する特徴を踏まえた次回作問時のアドバイスを生成する. 生成される作問アドバイスの一部を表 2 に示す. 現在, 作問アドバイスは全部で 60 種類登録されている.

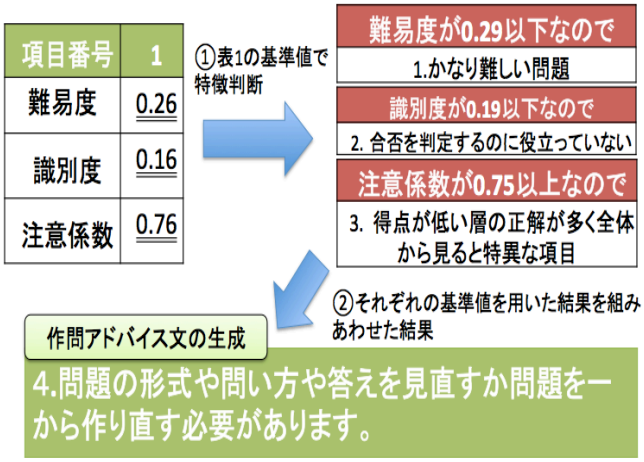


図 3 作問アドバイス生成の流れ

表 2 作問アドバイス生成例

条件	作問アドバイス例
難易度 0.3 未満 識別度 0.2 未満 注意係数 0.75 以上	問題の形式や問い方や答えを見直すか問題を一から作り直す必要が有ります.
難易度 0.3 未満 識別度 0.2 未満 注意係数 0.5 以上 0.75 未満	合否判定には機能していない問題であり, 問題の問い方や答えがあっているか注意してみる必要が有ります.
難易度 0.3 未満 識別度 0.4 以上 注意係数 0. 5 未満	難しい問題を作る際はこの問題をベースに問題を作るか, 修正すると次回のテストでも使えるでしょう.
難易度 0.4 以上及び 0.80 未満 識別度 0.4 以上 注意係数 0. 5 未満	最も理想的な問題であり, 次回以降のテストにおいても使えるでしょう. この問題をベースにすると理想的な問題ができそうです.

5. システム開発

上記の考え方に基づき項目分析支援システムを開発した. なお, 本研究では出題意図を考慮した分析結果やアドバイスの提示は未実装である. 開発言語は PHP, JavaScript, HTML, データベースには MySQL を用いた. 本システムで項目を分析する際に, テスト結果を正誤の 2 値データに置き換える. 置き換え方法としては, 各項目の正誤を 0 か 1（誤答：0, 正答：1）で表し, 項目反応データとし扱う. 項目反応データを本システムに入力し, 出力された分析結果の一例を図 4, 図 5 に示す. 本システムでは, まずテスト全体の結果が表示される（図 4）. そして, テスト全体の結果の下に, 良問と悪問の一覧と作問アドバイスが表示される（図 5）. 図 5 中の各項目番号のリンクをクリックすると同画面上にポップアップ表示され, 項目の詳細な特徴説明が表示される（図 6）. 同画面で各項目の特徴説明を表示することにより, 利用者のページ遷移の回数を少なくし, ページ読み込みの回数を最小限にしている.

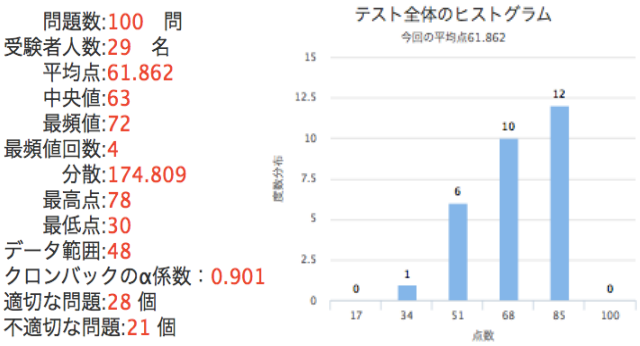
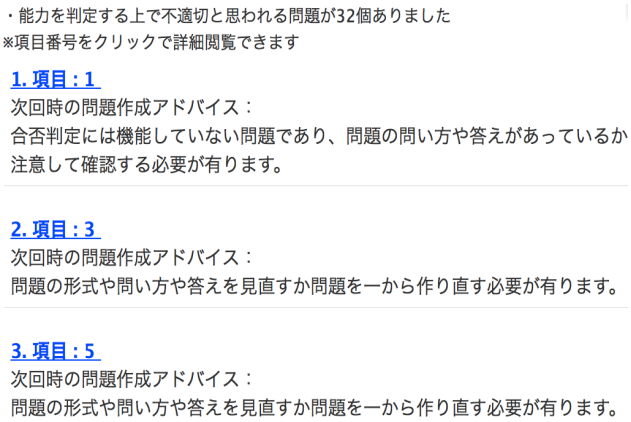


図 4 テスト全体の結果



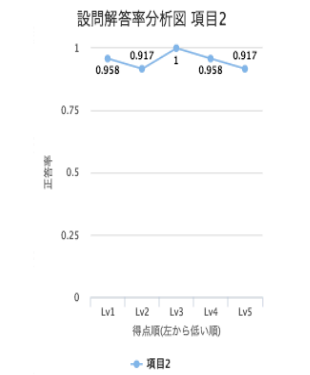
・能力を判定する上で適切と思われる問題が20個ありました

1. 項目 : 2
次回時の問題作成アドバイス：
数値上最も理想的な問題であり、次回以降のテストにおいても使えるでしょう。この問題をベースにすると理想的な問題ができそうです。

2. 項目 : 14
次回時の問題作成アドバイス：
数値上最も理想的な問題であり、次回以降のテストにおいても使えるでしょう。この問題をベースにすると理想的な問題ができそうです。

3. 項目 : 18
次回時の問題作成アドバイス：
数値上最も理想的な問題であり、次回以降のテストにおいても使えるでしょう。この問題をベースにすると理想的な問題ができそうです。

図 5　良問と悪問の一覧および作問アドバイス表示画面



グラフ：左側が低く、右側が高いと良い問題だといえるでしょう

難易度：　難易度が高い

識別度：　識別度が非常に低く、合否判定に全く適していない

注意係数：　注意係数が非常に高く、要注意な項目

閉じる

正解者数:114	不正解者数:6	正答率:95%

項目難易度	項目識別度	項目注意度
0.95	0.03	0.925

図 6　各項目の特徴説明

6.　システムの評価実験

6.1　実験の目的

本システムがテストの作問現場で活用でき，作問時に有効であるのかを検証するために，テスト理論の知見を有していないもりけん作問委員 6 名（2006 年のもりけん開催からの作問担当者）に，システムから提示される内容を閲覧してもらい，今回のテスト作成に活用することができたのかを実験する．

6.2　実験の方法・手順

実験の事前準備として，共同研究を行っているもりけん主催の岩手県盛岡市商工会議所から 2008 年度 2 級･3 級と 2009 年から 2015 年度 1 級･2 級･3 級（2013 年度は解答データなし）のもりけん解答結果を受け取った．本システムで分析するには解答結果を 2 値データに変換して分析する必要がある．そのため，システムに入力する解答結果のデータは，それぞれの年度と級ごとに excel で管理していた解答結果（合計 20 ファイル）を，著者が問題毎の正解・不正解の 2 値反応データに変換し，CSV ファイルとして用意した．変換したファイルをそれぞれ本システムに入力し，分析した結果をまとめたものをもりけん作問委員 6 名に提示した．また，分析結果に表示された難易度や識別度などの数値の意味や，設問解答率分析図の見方などを記述した補助資料を配布した．これらの分析結果と補助資料はもりけんの作問期間中に配布し，その後 3 週間に渡りもりけんの作問活動が行われた．その後，提示内容から把握できた過去問題の改善点や，過去問題を活用する上で参考になった部分などに関して半構造化インタビューを行い，本システムで分析した結果が作問現場で効果的に活用できたのかを考察する．

6.3　実験結果

インタビューの所用時間は一人あたり 5 分間で実施した．内容は表 3 の質問項目を基に行った．表 3①の質問項目では，6 名中分析結果を活用できたと答えた人が 3 名だけであった．活用できなかった 3 名の理由としては，忙しくて見る余裕がなかった，見方が分かりにくかったなどの理由が挙げられた．活用できた 3 名については表 3②以降の質問をした．表 3②の質問の結果，難易度，識別度や設問解答率分析図を参考にしたことが分かった．表 3③の質問では，難易度の高い問題や低い問題をテスト全体のバランスを考えて選択する際に参考にしたことや，難易度も高めで識別度も高い問題に関しては次回以降にも充分に参考にして活用できるので，難易度と識別度の高い問題を探す際に活用したなどの回答が多かった．他には，設問解答率分析図も合わせて見ることでより傾向が分かり，2 級問題で全体的に解けていない問題を 1 級に繰り上げることや，逆に 2 級問題で簡単だった問題を 3 級に繰り下げることの参考になったなどの回答があり，難易度・識別度・設問解答率分析図は充分な活用ができる可能性が示唆された．表 3④については，正答率や識別度の範囲を決めて検索などができるとより使いやすくなるという意見や，作成した問題やそのカテゴリなどの情報もシステムに入力して分析できるようになる

と，カテゴリの偏りなども調べることができ，より便利になるという意見を頂いた． また，分析して提示した問題数（1700 問）が多かったことから特定の難易度や識別度を満たす問題の検索や，右上がりの分析図などの検索ができる機能があると，より効率的にシステムが活用できることができるとの意見もあった．

表 3 インタビュー時に使用した提示項目

①. 提示内容を活用することができましたか？
②. 提示内容のどの項目（難易度，識別度など）を見ましたか？
③. 提示内容をどのように活用しましたか？
④. 提示内容で改善して欲しい箇所，欲しい情報や機能はありますか？

6.4 考察

本実験の結果から，分析結果には難易度・識別度・設問解答率分析図があれば，問題作成の参考になることが推察された．本実験での作問現場での活用に関しては，難易度のバランス調整などでは本システムの有効性が示唆された．しかし，難易度などの数値以外のカテゴリ情報まで考慮できていなかったため，来年度の活用に向けて本システムを改善する． また，インタビュー内容が終わった後に，もりけんの作問活動は 10 年ほど前から行われているが，本システムを導入した本年度の作問活動が一番楽で安心して試験作成ができたとの意見も頂くことができたため，本システムを活用することで，作問作業全体の負担を多少なりとも軽減ができたと考えられる．

7. まとめと今後の課題

7.1 まとめ

本稿では，作問者がテスト受検者の解答結果に基づいて次回作問時の改善点を把握することを目的とし，項目の分析支援システムの開発を行った．項目分析結果の，難易度と識別度と注意係数の結果から，それぞれの特徴を説明した文章を登録した DB と次回作問時のアドバイスルール DB の生成を行い，作問アドバイスを生成できる機能も開発した．開発した本システムの評価実験として，本システムがテストの作問現場で活用ができ，有効性があるのかを検証するための実験を行い，システムからの分析結果が作問作業において有効であることを示すことができた．

7.2 今後の課題

今後は，難易度や識別度検索など数値検索を導入していく．設問解答率分析図に関しても，分析図の特徴などで検索できるようにすることでより必要な情報を表示できるようなシステムに改善していく．また，今後の発展としては，分析結果を DB などに保存しテスト作成者向けに分析結果を蓄積し，過去の作成した問題の振り返りや自身が作成した問題の傾向などを再度確認できるような分析結果蓄積システムにする．この部分はこれからのシステムで重要な機能であり，システムを基に，作問時では気づかなかった点のフィードバックや新たな分析方法の参考となることが期待される． また，今回行った実験結果を踏まえてより利用者が使いやすい UI，UX を検討し，考慮したシステムに改良をする．

謝辞

調査の実施及び分析にご協力いただいた，岩手県盛岡市商工会議所の皆様，盛岡もの識り検定作問委員会の皆様に感謝いたします．

参 考 文 献

(1) 樋口三郎.『テストおよびアイテム分析 Web サービスの開発』.教育システム情報学会第 39 回全国大会講演論文集，p.377-378（2014）
(2) 熊谷龍一．初学者向けの項目反応理論分析プログラム EasyEstimation シリーズの開発．日本テスト学会誌，5，p.107-118（2009）
(3) 吉村宰：大学入試センターにおけるテストデータベースによる項目分析，植野真臣，永岡慶三（共編），e テスティング，培風館，pp.167-190（2009）
(4) 大友賢二.『言語テスト・データの新しい分析法 項目応答理論入門』．大修館書店，（1996）
(5) 山森光陽．前田啓朗（編）『英語教師のための教育データ分析入門』．東京:大修館，pp.4－12（2004）
(6) Nunnally,JumC.PsychometricTheory 2ndEdition . New York:McGraw-Hill Book Company, （1978）
(7) 佐藤隆博：S-P 表の入門（教育実践文庫 3），明治図書

出版社，（1985）

(8) 藤垣雅司, 藤垣康子, 中島光洋.「注意係数の規格化：S-P 表における反応パターンの指数について」. 日本科学教育学会，Vol.9，pp.260-261（1985）