様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　４月３０日現在

機関番号：３３１０９
研究種目：基盤研究（C）
研究期間：2010～2012
課題番号：２２５２０５９０
研究課題名（和文）　教員・学校間の協働作業による能力記述文との対応付けのある CAT 開発

研究課題名（英文）　Development of CAT associated with can-do statements in cooperation between teachers / schools

研究代表者
木村　哲夫（KIMURA TETSUO）
新潟青陵大学・看護福祉心理学部・教授
研究者番号：90249095

研究成果の概要（和文）：小規模であってもコンピュータ適応型テスト（CAT：受験者の解答が正解か不正解かによって、次の設問の難易度をコンピュータが調整して出題するテスト）を，複数の教員や学校が協働作業をすることによって、開発し実施できることを示すとともに、その結果を能力記述文と対応付けることを目指した。あわせて、学習管理システム（LMS：コンピュータ上で行われる学習を管理するシステム）で簡単に CAT を実施可能にするプログラムの開発を行い公開した。

研究成果の概要（英文）：This study demonstrated that small-scale in-house computer adaptive test (CAT: a form of computer-based test that adapts to the examinee's ability level) can be developed in cooperation between teachers / schools. This study also tried to associate test results with can-do statements. Besides, this study developed CAT programs that run on a learning management system (LMS: a software application for the management of learning on computer) and opened them to public.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	900,000	270,000	1,170,000
2011 年度	1,300,000	390,000	1,690,000
2012 年度	700,000	210,000	910,000
年度			
年度			
総　計	2,900,000	870,000	3,770,000

研究分野：人文学
科研費の分科・細目：言語学・外国語教育
キーワード：教育評価・測定、能力記述文、協働作業、LMS、CAT

１．研究開始当初の背景

(1) 大学に入学してくる学生の英語力は多様化しており、何らかの方法で習熟度別にクラスを配当し、いわゆる「リメディアル英語」の学習を必要とする学生を割り出す必要性が高まっていた。そのために利用可能な商用テストも開発されていたが、オープンソースのラーニング・マネージメントシステム（learning management system: LMS）のひとつ Moodle を利用して、十分な実用性と有効性のあるコンピュータを利用したテスト（CBT）実施が可能であることは、古典的テ

スト理論(classical test theory: CTT)の枠組みで Hinkelman & Grose(2004)で既に報告されていた。さらに、Kimura(2009)および木村(2009)では、項目応答理論(item response theory: IRT)およびニューラルテスト理論(neural test theory, NTT; Shojima, 2008)の枠組みで、同様の CBT 開発が可能であることを確認しており、データ数は少ないものの、商用テスト（TOEIC Bridge と CASEC）の結果とも高い相関（r_s=.80～.89）があり、その妥当性も検証されていた。

(2) 木村(2009)の CBT には、改善すべき点として次の 3 点があり、本研究を計画するに至った。

① 所要時間は、大学の授業 1 コマ 90 分にぎりぎり収まる長さであるが、より短時間で実施される方が望ましく、テストの信頼性と精度を下げずに項目数を減らし時間を短縮するために、受験者の応答によってテスト項目の難易度を調整するコンピュータ適応型英語テスト(CAT)の開発が望まれた。
② 結果は IRT によって標準化された間隔尺度上に、NTT によって標準化された順序尺度上に表すことが可能だが、それらの意味するところを解釈することが難しい。能力記述文との対応を示した商用テストとの相関分析により、間接的に現テスト結果の意味するところをある程度表現することは可能であるが、他のテスト結果との比較なしに、現テストの結果を直接何らかの能力記述文（Can-do statement: CDS）と対応付けることが望まれた。
③ 項目の種類として、語彙文法問題と会話と説明文によるリスニング問題しかないため、英語基礎力の測定を目的とするテストとしての構成概念的妥当性から、読解問題や作文問題を追加する必要もあった。

２．研究の目的

(1) 本研究の目標は、CDS との対応付けのある CAT 開発であるが、大規模に開発を行おうとするものではなく、教員・学校間が協同することで、小規模であっても CAT 開発できることをしめすことを目指した。研究のモットーは“CAT for everyone”である。本研究の目的は以下の 4 つからなる。

① CAT 実施の前提となるアイテムバンクを構築すること。
② LMS 上で CAT を実施するシステムを開発すること。
③ CAT の結果を何らかの CDS との対応付けを行うこと。
④ CAT の理論と実践について、より多くの教員・研究者に理解してもらうこと。

３．研究の方法

(1) 本研究で利用した項目は、すべて日本英語検定協会の許可を得て，英検準 1 級から 3 級の過去問題(2007～2008 年度)を利用した。項目の形式は、すべて 4 択の多肢選択問題で、項目の種類としては、次の 4 種類である。

① 文法語彙問題（vocabulary and grammar, Vgm）
② ダイアローグの聴解問題（listening comprehension with dialogues, Dlg）
③ モノローグの聴解問題（listening comprehension with monologues, Mlg）
④ 読解問題(reading comprehension, Rdg)

(2) 木村(2009)の CBT は、Vgm が 32 問、Dlg が 13 問、Mlg が 19 問、3 種類合計 64 問から構成されていた。これらの項目をアンカー項目として等化させるテストを複数用意し、教員・学校間の協働作業により事前テスト実施し、その結果を等化することでアイテムバンクを構築した。当初アイテムバンクは、項目の種類ごとに構築していたが、Dlg と Mlg は項目分析を進める中で、聴解力を測定する項目として 1 つのアイテムバンクに統合し、1 つの聴解問題（listening comprehension, Lng）のアイテムバンクとした。木村(2009)の CBT に欠けていた項目の種類として、Rdg を新たに追加する作業も平行して行った。

(3) LMS 上で CAT を実施するシステムとしては、Rasch モデルに基づき CAT を実装する BASIC で書かれたプログラム UCAT（Linacre, 1987）を元に、Moodle 上で実装できるように PHP でプログラムの書き換えを行った。当初 Moodle のバージョン 2.0 に合わせて開発を始めたが（Kimura & Ohnishi, 2011）、現在はバージョン 2.3 で開発を続けている(Kimura, Ohnishi, & Nagaoka, 2012)。

(4) テスト結果と CDS との対応付けのことを考えると、Rasch モデルや他の IRT のように連続尺度上で能力を評価する方法よりも、離散的な順序尺度上で能力を表現する潜在ランク理論（latent rank theory: LRT）に基づく項目分析と CAT の実装の方が段階評価に適しており、CDS との対応付けが容易に行えると考え、まだ提案されていなかった LRT に基づく CAT アルゴリズムの提案を行った。

(5) CDS については、新潟県内の 2 大学の 1 年生対象に開講された一般教養科目の英語 7 クラスの受講生合計 295 名（工学系・看護系・福祉心理系の学部学科）の協力を得て、STEP の英検 Can-do リストの CDS を使って、入学

時の英語力を自己評価してもらった。その上で、Lng と Rdg のテストの結果とこの CDS による自己評価の結果を比較し、テスト結果と CDS の対応付けについて考察を加えた。

(6) CAT の理論と実践について、より多くの教員・研究者に理解してもらうことを目的に、IRT と CAT については、Assessment Systems Corporation の協力を得て 2010 年 9 月に東京でワークショップを、小規模 CAT の開発のフレームワークについては、2012 年 8 月にシドニーで解された IACAT でシンポジウムを企画・開催した。

４．研究成果

(1) CAT 実施の前提となるアイテムバンクを、教員・学校間の協同作業によって構築することができることを示すとともに、シミュレーションと実テストの実施により、構築されたアイテムバンクの検証を行い、アイテムバンクの改善すべき点を明らかにできることを示した。

① 現在、アイテムバンク Vgm と Lng には、共通項目を含む複数のテスト結果から等化することによって、表 1・表２に示す数の項目が蓄積されている。参考として、Rasch モデルによる項目困難度の基本統計料を英検の級ごとにも示す。Lng は、もともと Dlg と Mlg という 2 つのアイテムバンクとして分析されてきたが、図 2 に示すように、この 2 種類（ダイアローグによる聴解問題とモノローグによる聴解問題）は、同一の傾向を示していることが確認されたので、1 つに統合された。

表１　アイテムバンク Vgm の Rasch モデルに基づく項目困難度基本統計量

	英検級				合計
	3 級	準2級	2 級	準1級	
N	49	67	69	73	258
M	-1.41	-0.47	0.52	1.57	0.19
SD	0.80	0.91	0.81	0.84	1.37
Max	0.43	2.00	3.19	3.50	3.50
Min	-3.38	-3.11	-0.64	-0.29	-3.38

表２　アイテムバンク Lng の Rasch モデルに基づく項目困難度基本統計量

	英検級				合計
	3 級	準2級	2 級	準1級	
N	80	75	109	44	308
M	-0.90	0.35	0.77	1.26	0.30
SD	1.33	1.05	1.11	1.42	1.43
Max	2.40	2.87	3.57	4.05	4.05
Min	-4.45	-2.44	-1.54	-2.25	-4.45

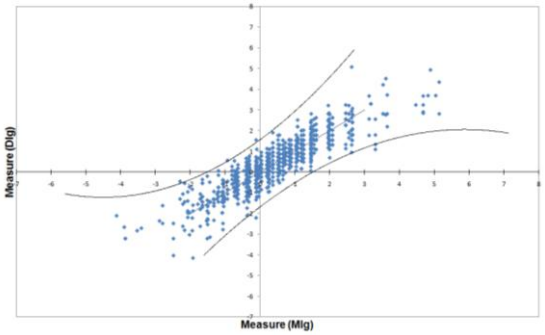

図２　Mlg *Measure* と Dlg *Measure* の統合

② 読解問題のアイテムバンク Rdg は、Vgm や Lng と異なり大問形式（１つの読解文に対して、2～5 の小問が配置されている形式）であり、多値データとして分析が行われた。蓄積された項目数は、大問単位でで 49 項目、小問単位で 175 項目になっている。

③ LRT に基づく CAT シミュレーションでは、項目の数が不足している潜在ランクの受験者に対しては、なかなか終了条件に達しないことが確認された（図 2 参照）。

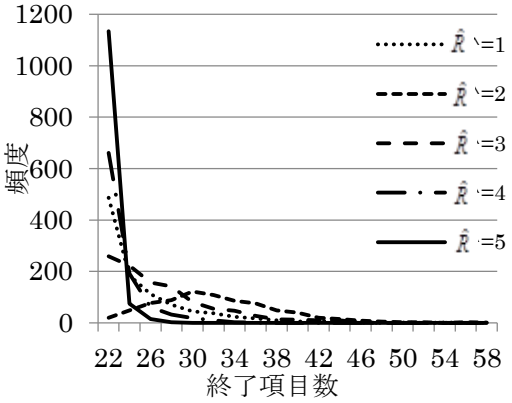

図２　$\hat{R}$ ごとの終了項目数

④ Rasch モデルの実テストでは、たとえば、項目困難度の分布状が図 3 のようなアイテムバンク(Vgm)で、CAT の終了条件を項目数 16 として、約 160 人の大学１年生に、に実施したところ、90％以上が S.E. 0.55logits 以下で終了していた。このアイテムバンクの項目使用頻度を受験者 100 人当たりで頻度を調べると、やさしいレベルの項目の頻度が高く不足気味であることが示された（図 4 参照）。

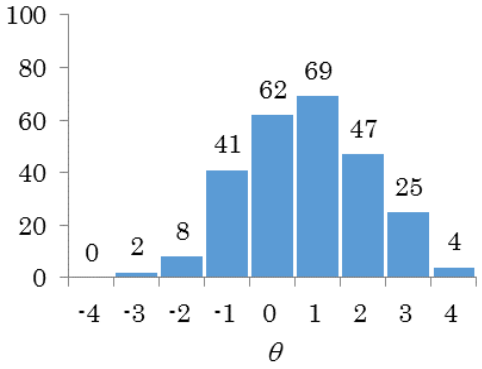

図 3　困難度ごとの項目数（Vgm）

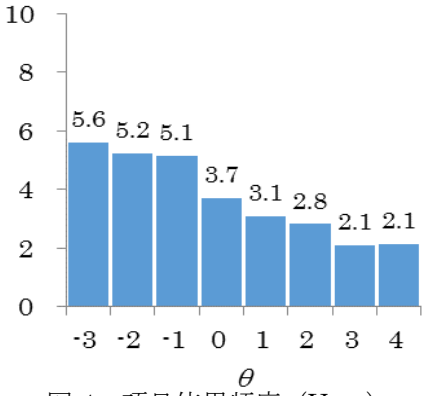

図 4　項目使用頻度（Vgm）

(2) LMS 上で CAT を実施するシステムとしては、次の 2 つのものを実装した。

① Moodle UCAT：Rasch モデルに基づく CAT 実行プログラム UCAT(Linacre, 1987) を元に、Moodle 上で同様の CAT を実装できるように PHP でプログラムの書き換えるとともに、目標正答確率を調整する仕組みを付け加え、β 版を国内外の学会で発表した。UCT の項目選択ルールでは、次に示す下限（lower limit: LL）と上限（upper limit: UL）の間の範囲から、ランダムに次の項目が選ばれて実施される。もし、この範囲の困難度の項目がアイテムバンクに存在しない場合は、この範囲から最も近い値の困難度を持つ項目が次の項目として選択される。m 項目終了後の LL は、

$$LL = \theta_m + \frac{R_{m-1} - \sum_{j=1}^{m} P_j(\theta_m)}{\sum_{j=1}^{m} P_j(\theta_m)\left(1 - P_j(\theta_m)\right)} \quad (1)$$

によって計算される。これは、m 項目が誤答であった場合の m 項目終了後暫定能力推定値である。UL は、

$$\boldsymbol{UL = LL + \frac{1}{\sum_{j=1}^{m} P_j(\theta_m)\left(1 - P_j(\theta_m)\right)}} \quad \boldsymbol{(2)}$$

によって計算される。Moodle UCAT では、(1)式に Logit Bias を加えて(3)式により LL を求めることで目標正答確率の調整を可能にした。Logit Bias に正の数値を入れると、選択される項目の困難度は高く、負の数値を入れると、選択される項目の困難度は低くなる。

$$LL_{biased} = LL + Logit\ Bias \quad (3)$$

Logit Bias を目標正答確率(p)を使って表現しなおすと、(4)式のようになる。

$$\boldsymbol{LL_{biased} = LL - \log_e\left(\frac{P}{1-P}\right)} \quad \boldsymbol{(4)}$$

(株)VERSION2 の管理するサーバー http://ucat23.moodle-ver2.jp/ にデモ環境を用意しプログラムを公開した。

② LRT に基づく CAT アルゴリズムとして、次のような仕組みを発表した。

初期能力推定値は RMP を使って、一様分布とする．すなわち、分析する潜在ランク数を Q とした場合、各潜在ランクに所属する確率を $1/Q$ とする。

最初の出題は 1 問ではなく、潜在ランクの中央値付近の問題を複数問出題するテストレット形式とする。

ランク数を Q とし、受験者 i が n 項目解答した時点の暫定 RMP を

$$\mathbf{p}_i^{(n)} = [p_{i1}^{(n)} \cdots p_{iQ}^{(n)}]' \quad (Q \times 1) \quad (5)$$

とする．ここで、$p_{iq}^{(n)}$ は、受験者 i の Rq に対する暫定的な所属確率である．
また、アイテムバンクにある j 番目の項目の IRP を

$$\boldsymbol{\upsilon}_j = [\upsilon_{j1} \cdots \upsilon_{jQ}]' \quad (Q \times 1) \quad (6)$$

とする。ここで，υ_{jq} は R_q に所属する受験者の項目 j に対する正答確率である。
さらに、IRP の差分ベクトル

$$\boldsymbol{\delta}_j = [\delta_{j1} \cdots \delta_{jQ-1}]' \quad ((Q-1) \times 1) \quad (7)$$

を計算する。ここで、

$$\delta_{jq} = \upsilon_{jq+1} - \upsilon_{jq} \quad (q = 1,2,\cdots,Q-1) \quad (8)$$

である。そして，$\mathbf{p}_i^{(n)}$ に対する識別度の高さを以下の式を用いて評価する。

$$\lambda_j^{(n)} = \frac{\sum_{q=1}^{Q-1} p_{iq}^{(n)} \delta_{jq} + \sum_{q=1}^{Q-1} p_{iq+1}^{(n)} \delta_{jq}}{2} \quad (9)$$

識別度の高いもの、すなわち、λ の値が最大のものから選択する方法も考えられるが、本研究の LRT-CAT の項目選択ルールとしては、アイテムバンクの中で λ の値最小のものから選択することを提案した。なぜなら、アイテムバンクがあまり大きくない場合、CAT 初期で識別度の高いものから実施すると、CAT 終期で RMP が収束し始めたときに、識別度の高いアイテム（局所的に（受験者の暫定ランクの付近で）急峻な IRP を持つ項目）がなく、かえって効率が悪くなることが懸念されたからである。CAT 初期の暫定 RMP はなだらかな形状であると考えられるので、識別度の低いものを実施し、識別度の高いものを CAT 終期に温存しておく方がよいと考えた。

アイテムバンクに蓄えられた項目数が多くなると、計算負荷がサーバーにかかり過ぎるので、本研究では IRP 指標 β が受験者 i の暫定の潜在ランクの推定値±1の項目に限定して λ の値を計算し、その中で最小となる項目を選択することとした(アルゴリズムの詳細については雑誌論文②を参照、プログラムについては、(株) e ラーニングサービスの管理するサーバー、http://moodle2x.info で公開)。

(3) CDS を使った自己評価への協力者のうち、リーディング (Rdg) とリスニング (Lng) のテストを受験したものについて，自己評価とこれらのテストの結果にどのようなずれが生じているか比較したところ、自己評価とテストの結果のランク数が完全に一致しているのは，リーディングで 35 人（30%）、リスニングで 40 人（29%）である。しかし、実際に行ったことがない事象についての記述も多い CDS に対して、必ず Yes/No で回答するのが難しいことも考慮に入れると、自己評価とテストの結果が少しだけずれることは自然なことと考えられる。自己評価とテストの結果のランク数が 1 だけ上下にずれている場合も、ほぼ一致しているとみなすと、リーディングで 66 人（57%）、リスニングで 85 人（61%）が自己評価とテストの結果がほぼ一致しており、約 6 割の学習者は自己評価とテストの結果にずれがほぼないことがわかる。

自己評価のランク数がテストの結果のランク数より 2 以上大きい場合を、過大評価（overestimate）、自己評価のランク数がテストの結果のランク数より 2 以上小さい場合を過小評価（underestimate）と定義することにした。今回はランク数を 5 とした分析なので、いずれの場合も最大 4 段階の差が生じうるが、いずれも 1～4%と極めて少ない。

リーディングとリスニングを比べると、過小評価となる者は、34 人(29%)対 24 人(17%)でリーディングの方が多く、過大評価となる者は、16 人（13%）対 30 人（22%）でリスニングの方が多い（図 5 と図 6 参照)。この割合の差は統計的にも有意な差である（χ^2= 5.84, df= 1, p = .016)。リスニングの方が、CDS に書かれている内容を実際に経験する機会が少ない（あるはまったくない）ことと、音声を伴わない文字媒体だけの状況での回答なので、「(やったことはないが）これくらいは，もしやればできるだろう」と判断してしまったためであろうことが推察される。

テストを実施するだけでなく、CDS を使って自己評価をさせることは、受験者が自分の能力を過大評価（あるいは過小評価）していることを気づかせるきっかけとすることもできる。しかし、テストによる能力評価と CDS による自己評価は、基本的に別次元のことを測定しているので、単純に両者をむすびつけることはできない。今後も CAT の結果を何らかの CDS との対応付けを行ことについて、さらに検討する価値は十分ある。

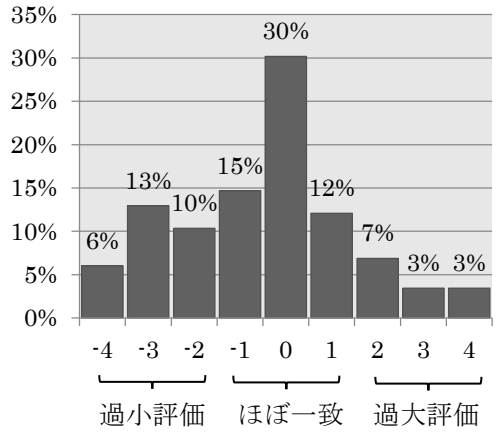

図 5　過大評価と過小評価の割合（Rdg）

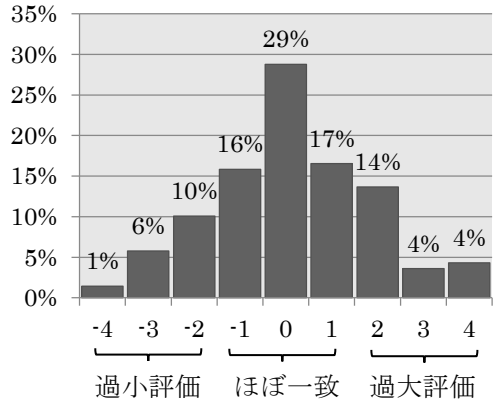

図 6　過大評価と過小評価の割合（Lng）

(4) Assessment Systems Corporation の協力を得て 2010 年 9 月に東京で行った IRT と CAT についてのワークショップは、国内外から約 40 人の参加者を得た。

2012 年 8 月にシドニーで解された IACAT で企画・開催したシンポジウムでは、Thompson & Weiss (2011)の案を元に、小規模 CAT の開発のフレームワークを示すとともに、2 つのアプローチ方法を示した。一つは、本研究成果の(2)のように、オープンソース LMS の Moodle に追加モジュールを開発し実装するもの、もう 1 つは、CAT 自体はオープンソースの R パッケージを元に開発された環境（Cambridge 大学 Psychometrics Center の開発した Concerto）で実装するものである。前者のアプローチの難点は、LMS 自体のプログラムのアップグレードに合わせて、開発した CAT の追加モジュールを修正しなければならなくなる点である。今後は、後者のアプローチを取り、LMS との間でデータ連携を行うことがよいことが示唆された。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計３件）

① 木村哲夫、コンピュータ適応型テストの心理学的側面：受験印象を改善するための目標正当確立を調整するシステムの実装、統計数理研究所共同研究リポート、295、2013、51-70
② 木村哲夫、永岡慶三、潜在ランク理論に基づくコンピュータアダプティブテストーアルゴリズムの提案と検証ー、日本テスト学会誌、査読有、8、2012、69-84
③ 小山由紀江、木村哲夫、Neural Test Theory を使った Can-do Statements の分析、統計数理研究所共同研究リポート、254、2011、59-77

〔学会発表〕（計２２件）

① 木村哲夫、永岡慶三、Moodle による小規模ＣＡＴ構築に向けて３：アイテムバンクの検証、日本教育工学会第 28 回全国大会、2012 年 9 月 15 日、長崎大学
② Kimura, T., Han, K. T., Kosinski, M., & Shojima, K., A framework and approaches to develop an in-house CAT with freeware and open source software. International Association of Computer Adaptive Test Conference 2012, 2012-08-13, Sydney Convention and Exhibition Centere, Australia
③ Kimura, T. & Nagaoka, K. , Can difficulty of items be guessed intelligently without degrading CAT results?, Pacific Rim Objective Measurement Symposium 2012, 2012-08-08, Fuyue Hotel, Jiaxing, China
④ Kimura, T., Ohnishi, A., & Nagaoka, K., Moodle UCAT: a computer-adaptive test module for Moodle based on the Rasch model., The 5th International Conference on Probabilistic Models for Measurement, 2011-01-15, University of Western Australia, Perth, Australia
⑤ Kimura, T. & Nagaoka, K. , Psychological aspects of CAT: How test-takers feel about CAT., International Association of Computer Adaptive Test Conference 2011, 2011-10-04, Asilomar Conference Center, California, USA
⑥ 木村哲夫、永岡慶三、Moodle による小規模ＣＡＴ構築に向けて 2：アイテムバンクの統合、日本教育工学会第 27 回全国大会、2011 年 9 月 19 日、首都大学東京
⑦ 木村哲夫、永岡慶三、潜在ランク理論に基づくコンピュータアダプティブテスト、日本テスト学会、2011 年 9 月 11 日、岡山大学
⑧ Kimura, T. & Nagaoka, K. , Reliability of Can-Do Statements about EFL Learners., Pacific Rim Objective Measurement Symposium 2011, 2011-07-14, The National Institute of Education, Nanyang Technological University, Singapore
⑨ 木村哲夫、教員・学校間で共有する英語基礎力測定のアイテムバンク、第 36 回全国英語教育学会大阪研究大会、2010 年 8 月 7 日、関西大学千里山キャンパス

６．研究組織

(1)研究代表者
木村　哲夫（KIMURA TETSUO）
新潟青陵大学・看護福祉心理学部・教授
研究者番号：90249095

(2)連携研究者
荘島　宏二郎（SHOJIMA KOJIRO）
大学入試センター・研究開発部・準教授
研究者番号：50360706

(3)研究協力者
永岡　慶三（NAGAOKA KEIZO）
早稲田大学・人間科学学術院・教授
研究者番号：90127382