研究報告

学力 / 非認知能力に対する対話・探究学習効果のデータサイエンス

― 全国学力・学習状況調査の分析を中心に ―

＊田端　健人，＊＊菅原　敏，＊＊板垣　将大，＊＊＊原田　信之，＊丸山　千佳子，
＊久保　順也，＊本図　愛実

Data Science for Effect Size of Dialogue and Inquiry-based Learning on Cognitive and Non-Cognitive Skills
- Focusing on the National Assessment of Academic Ability -

TABATA Taketo, SUGAWARA Satoshi, ITAGAKI Shota, HARADA Nobuyuki,
MARUYAMA Chikako, KUBO Junya and HONZU Manami

要　旨

本研究の目的は、対話・探究学習が学力と非認知能力に与える効果を測定し可視化することである。学習効果の可視化で著名なオーストラリアの教育学者ジョン・ハッティにならい、本研究でも可視化のために「効果量」を利用する。効果量は、およそ20年前に欧米で起きた「統計改革」により注目されるようになった数値であり、日本の教育界ではまだほとんど知られていない。そこで本研究では、効果量の定義や評価や利点を検討し、効果量を簡単に測定・可視化するウェブシステムを開発し、対話・探究学習の効果量を可視化した。このシステムの開発方針や開発方法については、本稿「補遺」で解説する。

利用データは、令和3年度全国学力・学習状況調査の貸与匿名データ約20万件である。非認知能力の定義を検討したうえで、児童生徒質問紙調査項目から非認知能力の指標となる「非認知『徳』尺度」と「対話・探究学習尺度」を独自開発した。効果の大きさを評価するため、家庭の社会経済的状況（SES）が学力に与える効果量を基準とし、これよりも対話・探究学習スコアの効果が大きければ、有意な効果があると評価することにした。結論としては、対話・探究学習は、学力に対してSESと同等かやや大きめの効果があり、非認知「徳」に対しては、SESとは比較にならない大きな効果があることがわかった。

Key words：学習効果の可視化、効果量、統計革命、基準値、社会経済的状況（SES）、ウェブシステム開発、エビデンス

＊　宮城教育大学教職大学院
＊＊　宮城教育大学教育学部
＊＊＊ 名古屋市立大学大学院

１．本研究のリサーチクエスチョン

現行の学習指導要領が提唱する「主体的・対話的で深い学び」の核心には対話的で探究的な学びがある。しかし、対話的で探究的な学びは、学力や非認知能力に対してどれほどの効果があるのだろうか。その効果は相対的に大きいのか小さいのか。この教育効果を数量的に測定できないか[1]。これが本研究のリサーチクエスチョンである。

２．先行研究からの示唆

メルボルン大学名誉教授ジョン・ハッティ（John Hattie）のメタ分析では、言語活動や探究学習の効果は高く評価されている。ハッティは効果量d=0.40を学校で1年間に達成できる平均的な値としている（cf., ハッティ, 2018, p.2）。

150項目のリスト上位から、言語活動と探究学習に関連する項目を紹介すると、表1になる[2]。なお略記号「ES」はEffect Sizeを意味し、効果量d値である。参考までに、学力と一定の相関があるとされる「SES（社会経済的状況）」や「家庭環境」の順位と効果量も併記する。

表１：言語活動と探究学習に関連する項目の効果量（ハッティのメタ分析より）

7位　学級での議論（ES=0.82）／17位　語彙力促進プログラム（ES=0.67）／18位　読書推進（ES=0.67）／21位　自己言語化（ES=0.64）／問題解決授業（ES=0.61）／26位　読解力の推進（ES=0.60）／44位　家庭環境（ES=0.52）／45位　社会経済的状況（ES=0.52）／91位　探究学習（ES=0.31）

探究学習の効果量は0.40に届かないものの、その他の言語活動や探究学習系の効果は、表1のようにかなり高く評価されている。

ただ、ハッティのメタ分析の対象となった900件を超える研究の著者は、ほぼすべてが日本人以外であり、データは海外の児童生徒のものと強く推測される[3]。

確かに、日本の学力調査研究にも、対話・探究学習と学力との相関係数を明らかにしたものがある。例えば、統計学の専門家である柴山直教授が協力し、新潟県教育委員会が実施した「平成16年度『全県学力調査』報告書」である。中学2年生239名のサンプルで、「児童生徒が、いろいろな考え方を発表したり、話し合ったりする授業を行っている」という質問項目について、Scheffeの事後検定によれば、「行っていない方が多い」群と「行っている方が多い群」との間に5%水準で有意な差が見いだされている（cf., 新潟県教育委員会, 2004, pp.263-264）。また総合的な学習で問題解決力や表現力をどの程度身につけているかを尋ねる5項目の合成変数「総合的な学習における態度得点」（クロンバックのα係数=0.75）では、この得点が高いほど学力が高いという傾向が示されている（cf., 新潟県教育委員会, 2004, p.251）。

同じく新潟県教育委員会「平成18年度『全県学力調査』報告書」では、対話と非認知能力との相関について、「『お互いに教え合う』、『発言や考えが大切にされる』、『話し合いができる』という学級の様子は、児童生徒の『自己効力感』、『成功体験』、『向上心・知的好奇心』、『根気強さ・ねばり強さ』、『人間関係』全てにおいて0.2以上の相関関係があった」（新潟県教育委員会, 2006, p.157）と報告されている。また「先生や友だちの話をしっかり聞き、集中して勉強に取り組む」と学力との相関については、中学校で相関係

1　第1執筆者は、対話的で探究的な学習の一つである「子どもの哲学p4c」の教育効果を、子どもの言葉の形態素解析により、言葉数や語彙数の伸びで可視化した(cf., 田端, 2021,第9・10章)。

2　ハッティ, 2018, pp.330-334の「到達度に及ぼす影響の一覧」より引用。

3　ハッティ, 2018, pp.274-329の「900超のメタ分析結果」一覧を参照し、日本人とおぼしき名前は、282番の「ヨシダ」、665番の「コバヤシ」、855番の「アキバ」の3名のみだった。

数が0.25、「授業中に自分の意見を進んで発表している」と学力との相関については、小学校で0.23と緩やかな相関が確認されている（cf.，新潟県教育委員会, 2006，p.160）。

しかし管見ながら日本の教育研究では、こうした効果に関する学力調査研究を見つけるのは骨の折れる作業であり、ハッティのように学習の効果量研究をメタ分析できるほどの量には到底達していない。第1・第7執筆者が所属する日本教育学会の学会誌『教育学研究』の投稿論文と特集論文を過去20年遡ったところ[4]、「効果量」やそれに関連する用語をタイトルとする論文は、わずか2本しかない[5]。その2本も効果量d値には触れていない。また第1・第3・第7執筆者が所属する日本教育方法学会の学会誌『教育方法学研究』の投稿論文を20年遡ったところ[6]、「効果量」をテーマとする論文は皆無であった。これら2つの学会誌ではそもそも数量的研究が極めて稀か皆無である。これら2誌ではハッティの翻訳書は書評も図書紹介もされていない。

ハッティがメタ分析している「効果量」は、実は世界的に見ても1990年代から2000年代初めにかけての「統計改革」（大久保・岡田，2012，p.5）をきっかけに注目されるようになった数値であり、2012年に出版された専門的調査研究によれば、日本では「効果量を重視する流れはほとんどない」（大久保・岡田, 2012，p.18）とされる。

統計改革については同書に詳しいが、例えばアメリカ心理学会（American Psychological Association）の出版マニュアルも第6版（2009年刊）からは、「帰無仮説検定は統計的分析のはじまりに過ぎず、効果量やCIなどを併せて記載することが結果を適切に報告するために必要である」[7]（cf.，大久保･岡田, 2012, p.11）と記されるようになった。心理学だけでなく、統計学の領域でも同様の動きがあり、アメリカ統計学会（American Statistical Association）も2016年に声明を発表し、「P値や統計的有意性は、効果の大きさや結果の重要性を意味しない」とし、P値以外のアプローチとして信頼区間やベイズ統計などを推奨している（cf., Wasserstein and Lazar, pp.2-3）。ハッティのメタ分析は、欧米でのこうした統計改革があってこそ可能になった研究である。

そこで本調査研究でもこの「効果量」を活用してみたい。その前に、日本の教育研究でほとんど知られていない「効果量」とは何か、その値をどう評価するか、どのような利点があるかを押さえておこう。

3.「効果量」について

3.1. 効果量とは何か？

「効果量」という用語は総称であり、70以上の種類がある（cf.，大久保・岡田，2012，p.46）。当然計算式も異なる。「効果量」は大きく2グループに分けることができる。一つは、「群間差についての効果量」であり、「2つの群の間にどの程度の違いがあるか」を表す「d族の（d family）効果量」である。もう一つは、「変数間の関係の大きさ」であり、「2つの変数間の関係がどの程度大きいか」を表す「r族の（r family）効果量」である（大久保・岡田，2012，pp.46-47）。前者は一般に効果量d、後者は相関係数rと呼ばれる。ハッティが扱う効果量はd値（族）である。

d族の効果量にしても複数の計算式がある。統計改革から20年程度経た現在でも、日本ではHedgesのgがdと称される混乱ぶりである（cf.，小林・濱田・水本, 2020，p.91）。

ハッティは次の計算式を記載している（ハッティ, 2012，p.337）。

効果の大きさ

$$= \frac{\textit{平均（ポストテスト）-平均（プリテスト）}}{\textit{広がり具合（標準偏差、または、sd）}}$$

ただこの計算式では、ポストテストとプリテストのどちらの標準偏差で割るのか、あるいは両者をプールした標準偏差で割るのか不明である。とはいえ、いずれの標準偏差をとるにせよ、この計算式でも効果量は計算できる。

4　『教育学研究』第89巻第1号(2022年3月) から第69巻第2号(2002年6月) までを調べた。

5　志水(2006) と川口(2006) の2本である。

6　『教育方法学研究』第47巻(2021年) から第27巻(2001年) までを調べた。

7　「CI」とは、「confidence interval(信頼区間)」の略称である。

いっそう正確を期すため、大久保・岡田（2021）が紹介するCohenのdとHedgesのgの計算式を示しておきたい。

グループ1の平均値をM_1、人数（サンプルサイズ）をn_1、分散をS_1^2（正確には「標本分散」）とし、グループ2の平均値をM_2、人数（サンプルサイズ）をn_2、分散をS_2^2とすると、Cohenのdの計算式は、以下となる（cf.，大久保・岡田，2012，p.55）。なお値がマイナスをとらないよう、絶対値とした。

$$d = \frac{|M_1 - M_2|}{S_p}$$

$$S_p = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}}$$

Cohenのdの計算式は、2群の平均値差を「プールした標準偏差（S_p）」で割っている。ちなみに、このdの式をコンピューターで計算しやすいよう、標準的なプログラム言語で記述すると、

d = abs(m1-m2)/sqrt((n1*s1^2+n2*s2^2)/(n1+n2))

となる。

次にHedgesのgの計算式は、「標本分散（S^2）」（大文字のS）ではなく、「不偏分散（s^2）」（小文字のs）を使い、次の式になる（cf., 大久保･岡田, 2012, p.56）。

$$g = \frac{|M_1 - M_2|}{s_p}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

これらの計算式からわかるように、効果量もサンプルサイズ（n_1とn_2）の影響を受けないわけではない。しかし影響を与えるのは、n_1とn_2の数値の「量」ではなく、両者の「割合」である。このことは、それぞれに次のような変数を代入してみるとよくわかる。コーエンのdのプールした標準偏差の式に、n_1=x、n_2=axと代入してみよう。つまり、グループ2の人数がグループ1の人数のa倍という想定である。するとプールした標準偏差は、

$$S_p = \sqrt{\frac{xS_1^2 + axS_2^2}{x + ax}}$$

となり、この式は、

$$S_p = \sqrt{\frac{(S_1^2 + aS_2^2)x}{(1 + a)\ x}}$$

と書き換えることができ、結果的に、

$$S_p = \sqrt{\frac{S_1^2 + aS_2^2}{1 + a}}$$

とx（サンプルサイズ）を消すことができ、n_1とn_2の比（1:a）のみが式に関与していることがわかる。

またdとgは、標本分散と不偏分散の違いや、プールした標準偏差の人数から2を引くか引かないかの違いはあるが、サンプル数が2群合わせて100程度を超えると、両者の違いはほとんどなくなる。

それぞれのプールした標準偏差S_pとs_pとの間には、

$$s_p = S_p\sqrt{\frac{n_1 + n_2}{n_1 + n_2 - 2}}$$

の関係があり、これに対応してdとgの間には

$$g = d\sqrt{\frac{n_1 + n_2 - 2}{n_1 + n_2}}$$

の関係がある（大久保・岡田，2012，p.57）。この関係から、サンプルサイズ（n）が大きくなるにつれ、定数「-2」の影響が小さくなりdとgが接近すること、そして定数「-2」が無視できるほど大きなサンプルサイズだと、ルート内の値が1に近似し、dとgが等しくなることが直感的に理解できる。ちなみに、n_1とn_2のサンプル数の合計が100になる場合のルート内の計算は98/100=0.98となり、その平方根を解くと0.99、つまりg=0.99dとなり、gとdは実質的に同じになる。

なお、dは「記述的な効果量」、gは「推測的な効果量」と解釈できる（大久保・岡田，2012，p.57）。

こうしたやや複雑な計算を簡略化するため、科研費基盤研究B「グローバル世界を視野とする学力・非認知能力の効果的学校モデル」（2020～22年度/課題番号:20H01667/研究代表:田端健人）の研究チームDS-EFA（Data Science of Education for All）は、2群の平均値、標準偏差、サンプル数を入力することで、P値、Cohenのd、Hedgesのg、CI、帰無仮説棄却の

有無、2群の正規分布曲線を出力するウェブアプリ「平均値差検定システム」を開発し、2022年5月より一般公開している（https://ds-efa.info/cohensd/）。このシステムではｄとｇの計算に上記の式を採用している。開発者は第2執筆者と第1執筆者である。本システムを学習効果の可視化に役立ててほしい。本システムの開発方針と開発方法については、本稿「補遺」で解説する。

検定ができる統計ソフトで比較的知名度が高い、js-STAR XR+[8]やHAD[9]と私たちが開発した「平均値差検定システム」とを比べてみると、js-STAR XR+やHADでは、私たちのシステムが出力する効果量のツールが装備されておらず、2群の分布曲線の描画機能がない。HADはヒストグラムを出力するが、私たちが開発した「平均値差検定システム」は、平均値と標準偏差から正規曲線を描き、2群の差をグラフで可視化する点に特長がある。またjs-STAR XR+やHADはSPSSと同様に、専門家でないと使いこなせないが、私たちのシステムは、平均値差に特化することで、シンプルであり、素人にもわかりやすく、教育現場で使いやすいものになっている。

さらに本研究チームは、すでに開発していた「全国学力・学習状況調査『平均ゾーンシステム』」（田端・丸山・本図，2022）のウェブアプリ（https://ds-efa.info/data_analysis/）に効果量ｄ値を追加した。このウェブアプリの開発者は第3執筆者である。ウェブアプリのもととなる理論面の開発者は第1・6・7執筆者である。効果量の追加により、個別の自治体や学校と全国平均との差を、分布曲線と効果量との2つで直感的かつ精度良く把握できる。こちらのシステムは、全国のトップ都道府県とボトム都道府県との分布曲線の間を「平均ゾーン」とみなす全く新しい発想で設計されており、現在のところ他に類を見ない。

これらのシステムは、自治体や学校の学力分析を支援する際に、たいへん重宝している。

3.2. 効果量の基準値

では、効果量ｄ値はどのように評価すればよいだろうか。これは「基準値」の問題である。効果量ｄの基準値（大きさの目安）には諸説ある。Cohenのｄの開発者ヤコブ・コーエン（Jacob Cohen）は表2の基準値を提唱する（cf.，Cohen，1988，p.82）。

表2：Cohenのｄの基準値

d = 0.8：効果大
d = 0.5：効果中
d = 0.2：効果小

この基準値には、「一定の有効性」があるとされ、「研究遂行前に、これは差がありそうだ、とか、厳しそうだ、という直感が働くことがあるが、その直感は概ねCohenの基準の大中小に沿うのではないか」と評価されている（村井・橋本，2018，p.123）。一方でコーエンの基準に否定的な見解もある（cf.，村井・橋本，2018，p.124）。学力/非認知能力のスコアを分析している第1執筆者には、コーエンの基準値は今のところ実用的にしっくりくることが多い。

ｄ値は平均値差に対してかなり感度が高い。このことはｄ値を、よく知られた相関係数ｒに変換すると直感的な理解が進む。コーエンによると、ｄからｒへの変換式は、

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

である（Cohen，1988，p.23）。エクセルなどで計算する場合は、「=d/sqrt（d^2+4）」と入力すれば良い。相関係数ｒは絶対値にして0から1までの値をとり、基準値としては一般的に表3とされる。

表3：相関係数ｒの基準値

0.7≦|r|≦1.0	：	強い相関
0.4≦|r|＜0.7	：	中程度の相関
0.2≦|r|＜0.4	：	弱い相関
0.0≦|r|＜0.2	：	相関なし

8　https://www.kisnet.or.jp/nappa/software/star/index.htm

9　https://norimune.net/had

次の表４は、ｄとｒの基準値の対応表である。

表４：ｄとｒ対応表

効果量dの基準	d	r	相関係数rの基準
	1.95	0.70	大
	0.90	0.41	中
大	0.80	0.37	
中	0.50	0.24	
	0.40	0.20	小
小	0.20	0.10	

d=0.50 は r=0.24 に相当し、弱い相関がある程度である。d 値で 0.50 もない場合は、差はないに等しいと評価してもよい。ハッティの d=0.40 は、r=0.20 であるから、相関があるといえる最低ラインである。

3.3. 効果量の利点

効果量ｄ値の利点を３点紹介しておきたい。

すでに言及したように、第１の利点として、統計改革の文脈において、効果量はサンプルサイズの影響をほとんど受けないことがある。逆に統計解析の主役であり続けている帰無仮説検定（＝統計的仮説検定）の問題点は、サンプルサイズが大きければ検出力が高まり、わずかの平均値差でも有意差ありになってしまうことにある。これは統計学の入門講義や入門書などで頻繁に言及される事項である（cf.，村井・橋本，2018，p.116）。ところがどういう訳か、ほとんどの研究が統計的有意性にだけ目を向け、サンプルサイズには無頓着である現状がなかなか変わらないという（cf.，村井・橋本，2018，p.117）。

帰無仮説検定とは、わかりやすく解説すると[10]、２群の平均値差が、帰無仮説を正しいとした場合に偶然に生じたものかそうでないかを確率的に判断する方法である。偶然に生じる差を小さな差、偶然には生じない差を大きな差とみなし、大きな差を「有意差」と呼ぶ。例えばコインを１回投げて表が出る確率は 1/2 で 0.50（50%）であり、これは偶然の結果である。コインを２回投げ２回とも表が出る確率は 1/2^2 で 0.25（25%）、感覚的にはこれも偶然に十分生じる確率である。しかし、４回（1/2^4=0.06）なり５回（1/2^5=0.03）連続で表となると、偶然とは思えなくなる。コインに仕掛けがしてあるだろうと疑いたくなる。帰無仮説検定では差の大小を判定するために、Ｐ値つまり Probability Value（確率値）を計算し、慣例でＰ値が 0.05 を下まわれば、有意水準 5% で有意差ありとする。この 0.05 という基準はあくまで慣例に過ぎないにもかかわらず、あたかもそれが絶対的客観的な基準と受け取られ、手続きとしては、「２群の平均値に差は無い」という帰無仮説（H_0）を立て、Ｐ値が 0.05 を下回れば帰無仮説を棄却し、対立仮説（H_1）「２群の平均値に差がある」を採択する、などのようにマニュアル化されてしまっている。

帰無仮説検定の難点は、全国学力・学習状況調査の都道府県のサンプルサイズになると常に顕在化し、ほとんど意味のない平均値差でもＰ値は 0.05 を下回り、有意水準 5% で有意差ありとなる。令和３年度全国学力・学習状況調査の小６国語の全国平均正答率を利用し、先に紹介した「平均値差検定システム」でシミュレーションしてみてほしい。

令和３年度小６国語の全国（公立）平均正答数は 9.1 問、標準偏差は 3.1、児童数は 993,975 名である。シミュレーションとして、宮城県と同じ児童数 18,096 名でかつ同じ標準偏差、平均正答数が 100 分の１（=0.091: 正答率にして１ポイント）低い 9.009 という架空の自治体を想定してみよう。100 点満点のテストの平均で１点差であるから、現実的には差は無いに等しい。システムが出力する分布曲線もほとんど重なり合って区別がつかない（図１のグラフ参照）。しかし、この僅差でさえ、サンプルサイズがこれだけ大きければ、有意差ありとなる（図１の「Ｐ値」や「有意水準 5%」の欄の評価を参照）。

このようにたとえ仮説検定で有意差があるからといって、正答率１ポイント差を意味のある差とするのは全く現実的ではない。ちなみにサンプルサイズを 8,900 名にすると、p=0.05 となり有意差なしになる（図２参照）。このように全国学力・学習状況調査では、Ｐ値は全く役に立たない。

ここで効果量の重要性が浮上する。システム出力の「Cohen の d」をみると、図１でも図２でも、効果量は、サンプルサイズに左右されず 0.03 と非常に小さいこ

10　さらに専門的な解説については、大久保・岡田(2012) の「2.2　帰無仮説検定の論理」を参照されたい。

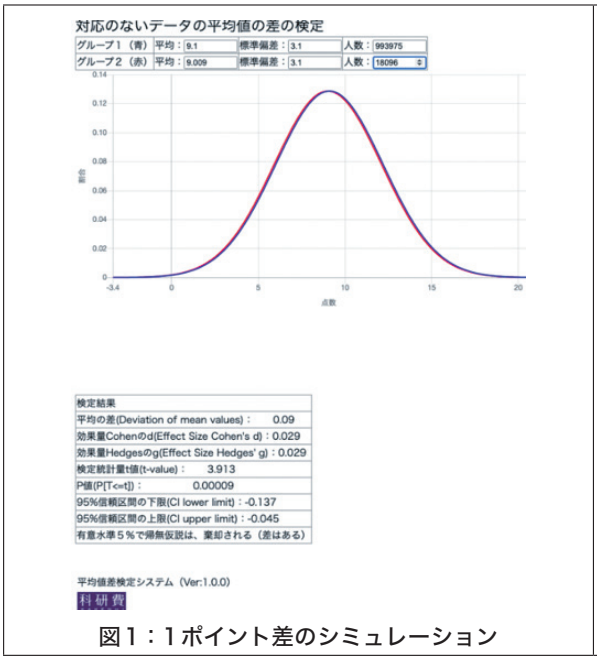


図１：１ポイント差のシミュレーション

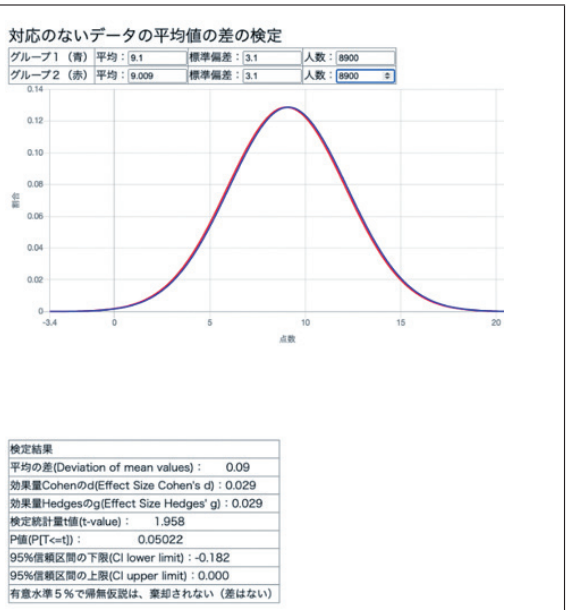


図２：図１のサンプルサイズを小さくした場合

とがわかる。この学力スコアの平均値差の評価としては、P値より Cohen の d値の方が役に立つ。

第２の利点として、効果量は標準化されており、元データの観測変数の単位に依存しない。測定方法の違いを超え、時間差を超え、グループの違いを超えて、一つの尺度で平均値差の大きさを比較できる利点が効果量にはある（cf.，ハッティ，2012，p.4）。教育研究の領域において効果量に着眼したことは、ハッティの先見の明である。それゆえ、様々な学習法の効果やグループの平均値差や年度ごとの経年変化が、効果量d値で比較可能になる。全国学力・学習状況調査は古典的テスト理論で設計されているため、経年比較できないが、効果量を計算すれば実用レベルで便宜的に経年比較できるようになる。

第３に、効果量は直感的に理解しやすく、統計を専門としない教育研究者や実践者にもわかりやすい。d=1.0は、標準偏差１シグマ（σ）分の差になる。よく知られた「偏差値」は、平均50、標準偏差10の正規分布であり、標準偏差１シグマは偏差値10になる。d=0.5ならば、偏差値50と55の差に等しい。この程度の差を有意な差と評価する。これは私たちの一般的な数値感とも整合的であるだろう。

４．研究デザイン

そこで日本の児童生徒のデータから、対話的で探究的な学習の効果を示す新たなエビデンスを得てみよう。

次の研究デザインにより、対話や探究学習の効果量を測定・可視化することを試みる。

利用データには、令和３年度全国学力・学習状況調査の匿名貸与データ、小学６年生と中学３年生各約10万名、計約20万名を利用する[11]。

学力スコアには、国語と算数・数学の正答数を利用する。国語の正答数を「Jpn」、算数・数学の正答数を「Math」と略記する。

SES の代替指標には、児童質問紙項目（22）の蔵書数を利用する。「SES」と略。

児童生徒質問紙項目から、非認知能力に関わる合成変数（「Virtus」と略）と、対話や探究学習に関わる合成変数（「Dial_Inq」と略）を新規開発する。開発者は、第１・５・６・７執筆者である。

SES と教科の学力との相関については、多くのエビデンスがあるため、その相関や効果量の程度を基準とし、Dial_Inq と学力との効果量、また Dial_Inq と Virtus との効果量を評価する。Dial_Inq に SES と同

11　令和３年の匿名データにつき、発表者は文部科学省の貸与制度を利用し、2022年２月から貸与を受けている。

等かそれ以上の相関や効果量があれば、Dial_Inq には有意な教育効果があると評価する。

５．各尺度と利用データの解説

5.1.社会経済的状況（SES）尺度

児童生徒の家庭の蔵書数は、社会経済的状況の代替尺度とされる指標である。令和３年度全国学力・学習状況調査では、蔵書数の質問項目があるため、本研究ではそれを SES の代替変数とする。なお、児童生徒が回答した家庭の蔵書数に、SES の代替変数としてどれほどの精度があるか、本研究チームで目下検証中である。国際的に精度の高い調査でも利用されるこの変数には、一定の信頼度があると仮定して議論を進める。

令和３年度児童質問紙調査では、SES の質問項目の回答は６件法で、「１ ０～10冊」「２ 11～25冊」「３ 26～100冊」「４ 101～200冊」「５ 201～500冊」「６ 501冊以上」となっている。児童がイメージしやすいように、次のようなイラストも添えられている[12]。

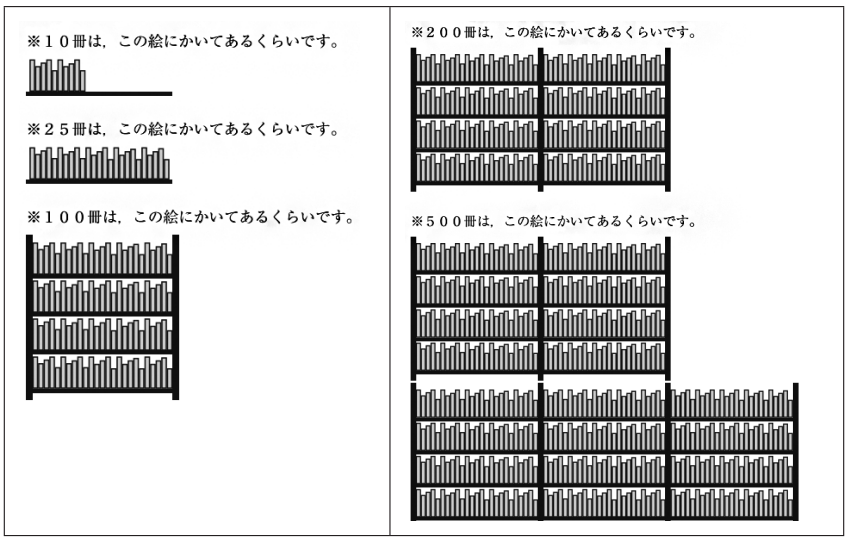


図３：平成３年度、全国学力・学習状況調査の児童質問紙調査の蔵書数質問項目

文部科学省は SES の代替変数となる質問（22）と学力との相関係数も発表している[13]。それを一覧化すると表５上段になる。文部科学省はそれらの相関係数を負の値で発表しているが、表５では絶対値に変更した。また Cohen の効果量ｄ値は、先述のように相関係数ｒに変換できるため、表５下段に効果量ｄ値を併記した。

表５：平成３年度調査、学力と SES の相関係数と効果量

	小6国語	小6算数	中3国語	中3数学
学力とSESの相関係数r	0.233	0.251	0.220	0.203
学力に対するSESの効果量d	0.478	0.518	0.452	0.414

12　国立教育政策研究所のウェブサイトで公開されている。以下の PDF の p.13より転載。
https://www.nier.go.jp/21chousa/pdf/21shitumonshi_shou_jidou.pdf

13　相関係数の調査結果は、国立教育政策研究所のウェブサイトから、「全国学力・学習状況調査」>「令和３年度報告書・調査結果資料」より、小６の結果分析は、https://www.nier.go.jp/21chousakekkahoukoku/factsheet/primary.html
中３の結果分析は、https://www.nier.go.jp/21chousakekkahoukoku/factsheet/middle.html
のページに進み、ページ末尾の「相関係数、クロス集計表」の「相関係数(児童生徒質問紙-教科)全国【表】」をクリックすると、自動ダウンロードされる。

先の表1に記したように、ハッティによるとSESの平均的効果量dは0.52、相関係数rにして0.25である。これと比べ、表5のSESの相関係数と効果量は若干低めであるが、おおよそ整合的である。

5.2.「非認知『徳（Virtus）』尺度」の開発

非認知能力（Non-Cognitive Skills）については、ジェームズ・ヘックマン（James Heckman）の提唱以来、日本でも注目が集まっている。「cognitive skills」と「non-cognitive skills」という英語をGoogle Books Ngram Viewer[14]で検索すると（2022年12月12日アクセス）、図4のように前者に比べ後者の使用頻度は圧倒的に低く、前者が1960年ごろから急速に使用されてきたのに対し、後者は、2010年ごろから使用頻度が上昇し始めた新しい言葉であることがわかる。

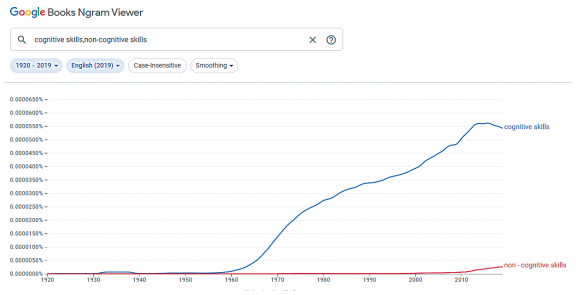


図4　英語圏での「cognitive skills」と「non-cognitive skills」の使用頻度の比較

非認知能力は、その名称から「認知能力以外の能力」と解釈されることがあるが、少なくともヘックマンの主要な論文を見る限り、ヘックマンがそのような定義をしたことはない（cf., Heckman and Rubinstein, 2001; Heckman et.al, 2006; Heckman et.al, 2013; Cuha and Heckman, 2006）。そもそもヘックマンは、認知能力とは何か、非認知能力とは何かという哲学的議論には立ち入っていない。

認知能力について、彼は「cognition, as measured by IQ」（Heckman et.al, 2013, p.2066）とか、「We use the Stanford-Binet Intelligence Test … as our measure of cognition」（ibid,p.2066）と述べている。つまりヘックマンのいう認知能力は、IQテストあるいは「standardized achievement and ability tests（標準化された到達度・能力テスト）」（Heckman and Rubinstein,2001,p.145）で測定される能力である。

同様にヘックマンのいう非認知能力は、心理学者が開発した「the Pupil Behavior Inventory（PBI）」や「the Ypsilanti Rating Scale（YRS）」で測定される「パーソナリティ能力」である（Heckman et.al,2013,p.2067）。ヘックマンらが分析し有名になったペリー計画が実施された頃にはまだ開発されていなかったが、「ビッグ・ファイブ（the Big Five personality inventory）」もこれらPBIやYRSと対応可能である[15]。

OECDも、非認知能力の別名である「社会情動的スキル」を具体化するためにビッグ・ファイブを引いている（cf., 経済協力開発機構, 2019, p.53）。ビッグ・ファイブの構成要素は、「開放性（Openness）」「誠実性（Conscientiousness）」「外交性（Extraversion）」「協調性（Agreeableness）」「情緒安定性（Neuroticism）」である（cf., ibid, p.53）。これらの要素で測定されるのが、ヘックマンやOECDのいう非認知能力である。

以上の検討からわかるのは、非認知能力は、学力と同じく、それが何であるか哲学的に解明された実態としての能力（真値）ではなく、パーソナリティテストで測定された数値（測定値）である。古典的テスト理論風に言えば、ヘックマンが有名にした「非認知能力」は、パーソナリティテストの測定値から測定誤差を差し引いた「理論的なあるいは概念上の値」（柴山, 2000, p.455）である。

非認知能力への注目の影響からか、全国学力・学習状況調査の児童生徒質問紙にも、パーソナリティテストと対応可能な質問項目が盛り込まれている。令和

14　https://books.google.com/ngrams

15　Heckman et.al(2013) のオンライン・アペンディクスの「Table C.1」や「Table D.1」では、PBIとYRSとビッグ・ファイブとの対応関係が一覧化されている。http://jenni.uchicago.edu/Perry/factor/0_PerryFactorWebAppendix_2012-11-24a_sjs.pdf

なおこの対応一覧の作成には、「グリット（やり抜く力）」で著名なアンジェラ・ダックワースが協力している。

３年度調査でみると、小６中３調査の質問番号６～16は、文部科学省の分類では「挑戦心、達成感、規範意識、自己有用感等」とまとめられている[16]。これらは非認知能力に対応すると解釈できる。

そこで執筆者らの研究チームは、全国学力・学習状況調査のこれらの質問項目と、関連する１項目を加えた12項目で、非認知能力の合成変数を開発してみた。その質問項目は表６のとおりである。試みにヘックマンの対応表も参照しビッグ・ファイブと対応づけてみた。

表６:非認知「徳」尺度の項目

（６）	自分には，よいところがあると思いますか	N
（７）	将来の夢や目標を持っていますか	C/N
（８）	自分でやると決めたことは，やり遂げるようにしていますか	C
（９）	難しいことでも，失敗を恐れないで挑戦していますか	E/O
（１０）	人が困っているときは，進んで助けていますか	E/A
（１１）	いじめは，どんな理由があってもいけないことだと思いますか	E/A/C
（１２）	人の役に立つ人間になりたいと思いますか	E/A
（１３）	学校に行くのは楽しいと思いますか	N
（１４）	自分の思っていることや感じていることをきちんと言葉で表すことができますか	E/C/A
（１５）	自分と違う意見について考えるのは楽しいと思いますか	N/A/O
（１６）	友達と協力するのは楽しいと思いますか	N/A
（２５）	地域や社会をよくするために何をすべきかを考えることがありますか	E/O

Notes: ビッグ・ファイブとの対応については、Openness(O), Conscientiousness(C), Extraversion(E), Agreeableness(A), Neurotixism(N)とした。

令和３年度全国学力・学習状況調査の匿名データの児童生徒質問紙調査結果から、この合成変数の信頼性係数（クロンバックのα）をIBM社の統計ソフトSPSSで算出したところ、小６で0.79、中３で0.80となり、信頼できる値となった。またSPSSの「項目を削除したときの尺度」の出力でも、これら12項目のうちどれか１つでも削除すると信頼性係数が低下することを確認した。

なお、データ入力に際しては、無回答「０」とその他「99」と不明な値「５」をエクセルで空欄に置換し、元々の欠測値である空欄と共に、SPSSにて空欄のあるデータを削除している。

独自開発の合成変数であるため、尺度名について研究チームで検討し、第５執筆者が「非認知『徳』尺度（スコア）」と命名した。学校は子どもの「知徳体」の成長に責任をもつ場所という認識が学校現場にあり、学力テストで測定される「知」と、体力テストで測定される「体」に対し、この合成尺度はそれらのテストでは直接測定されない「徳」の要素を測定すると考えたからである[17]。

5.3.「対話・探究学習（Dial_Inq）尺度」の開発

全国学力・学習状況調査には、児童生徒が学校で対話や探究学習にどのくらい取り組んでいるかを測定する質問項目も盛り込まれている。児童生徒の自己評価として、児童生徒がどのくらい対話や探究学習に取り組んできたかを、これらの項目から私たちは知ることができる。そこで、本研究チームは、関連する複数の項目を抽出し、「対話・探究学習尺度（スコア）」を独自開発した。

抽出した項目は、表７の通りである。表７は小６児童への質問紙項目であるが、中３生徒にも同じ番号で同じ質問がなされており、違いとしては、「５年生までに受けた授業」の部分が、「１、２年生の時に受けた

16　文部科学省「令和３年度 全国学力・学習状況調査 報告書―質問紙調査―」の「目次」の次のページ（ページ数なし）に分類表が記載されている。https://www.nier.go.jp/21chousakekkahoukoku/report/data/21qn.pdf

17　さらに広い文脈として、中国に由来し日本の文化として受け継がれた「九徳」と、古代ギリシアのソクラテスのいう「徳（アレテー）」や古代ローマで重視された「徳（ヴィルトゥース）」という東西の歴史も意識して命名した。

授業」という表記になっているだけである。

文部科学省の分類では、質問番号31～38は「主体的・対話的で深い学びの視点からの授業改善に関する取組状況」に、また質問番号39～42は「総合的な学習の時間、学級活動、特別の教科道徳」にカテゴライズされている[18]。本研究チームは、質問内容からして、これらは対話と探究学習に関わるまとまりのある項目と判断した。

なお、今回開発した2つの合成変数の項目間に重複がないかについて、統計的な検証はしていない。利用した項目の回答変数の因子分析をすれば、これらの項目をいくつの因子に分けることができるか統計的に明らかにできるし、2つの因子に分けた場合の項目の寄与度を測定することもできる。因子分析をすると、「非認知『徳』尺度」の項目でも、「対話・探究学習尺度」の方がふさわしいという結果になるかもしれない。しかし、本稿で測定したいのは、「対話・探究学習が学力や非認知能力にどの程度影響を与えるか」であるため、普段の学校生活全般で培われるパーソナリティに関する質問項目と、授業等での学習状況における対話や探究活動に関する質問項目とにまとめるだけで十分であり、因子分析の必要はないと判断した。

表7の右2列は文部科学省が公開している小6学力スコアとの相関係数である。相関係数0.20以上の相関があるセルをピンクでハイライトした。いくつかの項目で学力と弱い相関が確認されるため、「対話・探究学習尺度」は学力とある程度相関すると予想できる。問題は、その相関がSESを目安として、同等かそれ以上と評価できるか否かである。

ちなみに、「子どもの『徳』尺度」の質問項目ではいずれの項目も、文部科学省公開データでは学力との相関係数で0.20以上はない。

合成変数「対話・探究学習尺度（Dial_Inq）」の信頼性係数αをSPSSで確認したところ、小6で0.869、中3で0.872と高い信頼性が得られた。またどの項目を削除しても相関係数は低下することも確認した。

表7：対話・探究学習尺度の項目

質問番号	質問内容	相関係数 国語	相関係数 算数
（31）	5年生までに受けた授業で，学級の友達との間で話し合う活動では，話し合う内容を理解して，相手の考えを最後まで聞き，友達の考え（自分と同じところや違うところ）を受け止めて自分の考えをしっかり伝えていましたか	0.205	0.199
（32）	5年生までに受けた授業で，自分の考えを発表する機会では，自分の考えがうまく伝わるよう，資料や文章，話の組立てなどを工夫して発表していましたか	0.237	0.236
（33）	5年生までに受けた授業では，課題の解決に向けて，自分で考え，自分から取り組んでいましたか	0.246	0.252
（34）	5年生までに受けた授業では，各教科などで学んだことを生かしながら，自分の考えをまとめたり，思いや考えをもとに新しいものを作り出したりする活動を行っていましたか	0.181	0.173
（35）	5年生までに受けた授業は，自分にあった教え方，教材，学習時間などになっていましたか	0.152	0.155
（36）	友達と話し合うとき，友達の話や意見を最後まで聞くことができていますか	0.091	0.071
（37）	学級の友達との間で話し合う活動を通じて，自分の考えを深めたり，広げたりすることができていますか	0.191	0.185
（38）	学習した内容について，分かった点や，よく分からなかった点を見直し，次の学習につなげることができていますか	0.222	0.228
（39）	総合的な学習の時間では，自分で課題を立てて情報を集め整理して，調べたことを発表するなどの学習活動に取り組んでいますか	0.224	0.200
（40）	あなたの学級では，学級生活をよりよくするために学級会で話し合い，互いの意見のよさを生かして解決方法を決めていますか	0.140	0.120
（41）	学級活動における学級での話合いを生かして，今，自分が努力すべきことを決めて取り組んでいますか	0.088	0.077
（42）	道徳の授業では，自分の考えを深めたり，学級やグループで話し合ったりする活動に取り組んでいますか	0.139	0.115

18 注12を参照。

６. 学力スコアと SES と対話・探究学習の相関係数と効果量

6.1. データセットの紹介

次の作業として、令和３年度全国学力・学習状況調査の貸与匿名データでデータセットを作成し、欠測値を SPSS により削除した後、統計ソフト R[19]にて相関関係を可視化した。

なお、文部科学省の調査の回答では「1 当てはまる」「2 どちらかといえば、当てはまる」「3 どちらかといえば、当てはまらない」「4 当てはまらない」となっており、当てはまるほど合成変数の値が小さくなるため、SPSS にて得点を逆転した。

小６のデータセット（暫定名「dat」）の構造を示すために、R でデータの冒頭と行列数を表示した（表８）。

表８:統計ソフト R による可視化のためのデータセット紹介

```
> dat<-read.csv(file.choose(),header=T)
> head(dat)
         ID  EL_Jpn  EL_Math  SES  Virtus  Dial_Inq
1 ST000001       4        6    2      32        32
2 ST000002      12       14    3      42        41
3 ST000003       6       13    3      33        39
4 ST000004       7       10    2      39        46
5 ST000005      11       15    4      42        45
6 ST000006       2        8    2      34        30
> dim(dat)
[1] 96229     6
```

dim関数の結果に示されているように、小６のサンプル数は96,229名である。６列のヘッダーは、head関数の結果の１行目にある通りである。中３のサンプル数は85,338名であった。

6.2. 相関係数、ヒストグラム、分布図

R の psych パッケージ、pairs.panels関数でそれぞれの相関係数、ヒストグラム、分布図を一覧で可視化した（図５）。左が小６、右が中３である。

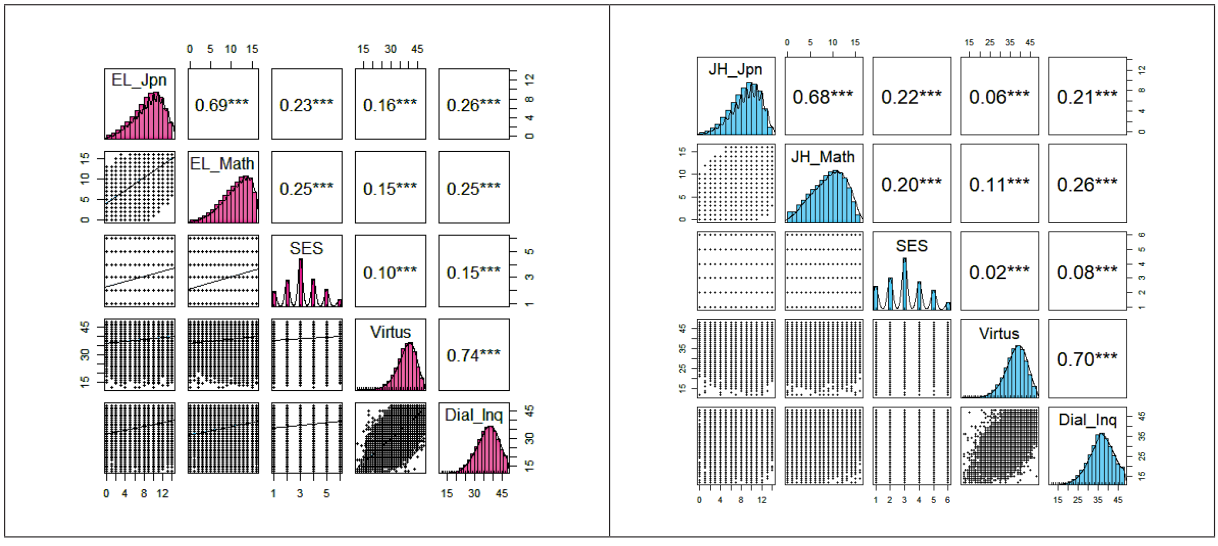


図５：令和３年度の小６（左）と中３（右）の各スコア間の相関係数

19 https://www.r-project.org/

アスタリスク *** は、p<0.001 で有意であることを意味する。しかしサンプル数が多いため、先述の帰無仮説検定のサンプル数問題の通り、この p 値は全く参考にならない。

国語スコアと算数・数学スコアの相関係数が0.69（小6）と0.68（中3）と出ており、強い相関が認められる。管見ながら、教科間の相関係数を明らかにしたこと自体、本研究独自の成果ではないだろうか。

図5の SES と学力との相関係数を見ると、小6国語と算数、中3国語と数学の値がすべて、文部科学省発表の表5の値と一致している。本研究のデータ処理と計算に間違いがなかった証左でもある。国語と算数・数学のいずれもで、SES との相関係数が、小6より中3で小さくなっている。小6よりも中3の方が、学力に対する SES の影響力が低下しているとの仮説も浮上する。

Dial_Inq と学力との相関係数は、小6中3の2教科ともに SES と学力との相関係数に匹敵する結果となっている。SES と学力の相関係数の平均（小中2教科）は0.23、d値にして0.46となり中程度に迫る効果量である。Dial_Inq と学力との相関係数平均は0.25、d=0.51 となり、SES の効果量よりわずかに大きい。ここから、本研究のリサーチクエスチョンへの一つのエビデンスが得られる。すなわち、対話・探究学習は学力に対して、SES と同等かときにそれ以上の効果をもたらす。

注目したいのは Virtus と Dial_Inq の相関である。一般に非認知能力に関する話題では、学力との相関関係に注目が集まりがちである。しかし、本分析から、学力と非認知能力とには直接的な相関がないことが判明する。これは、文部科学省が公開している、合成前の個々の質問項目と学力スコアとの相関のなさからも示唆されていた結果である。

本研究の独自性は、非認知能力の1指標である「非認知『徳』尺度」を、学力とは独立に扱い、学力と並ぶ重要な能力としてとらえる点にある。つまり非認知能力は、学力と相関しようとしまいと、それだけで価値ある能力であり、学校が育成すべき重要な能力であるというとらえである。学校は「知徳体」いずれの育成にも責任をもつという、この尺度の開発理念の通りである。

この分析結果から新たに判明するのは、Virtus と SES との相関係数が、小6中3とも0.20に届かず、非認知能力と SES との間には相関がない、ということである。SES が高いからといって Virtus が高いわけではないし、SES が低いからといって Virtus が低いわけでもない。非認知能力と SES とは相互に独立性が高い。

対して、Virtus と Dial_Inq の相関係数は小6で0.74、中3で0.70と強い相関が認められる。効果量dにするとそれぞれ、2.23と1.99となり、非常に大きな効果になる。対話・探究学習をしていると感じる児童ほど、非認知能力も高いという結果である。これは、子どもの非認知「徳」を育成するには、対話・探究学習が非常に効果的であることを示唆している。

6.3. 学力や非認知能力に対する対話・探究学習の効果の可視化

SES を基準とした対話・探究学習の効果を、箱ひげ図で比較することで、少し異なる角度から直感に訴えてみたい。

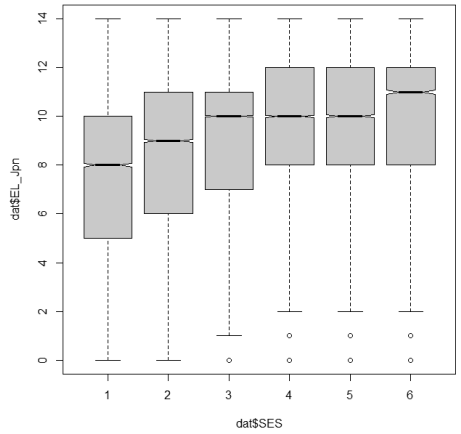
国語や算数・数学の学力を縦軸、SES や Dial_Inq を横軸とした箱ひげ図、ならびに Virtus を縦軸、Dial_Inq を横軸とした箱ひげ図を以下に示す。図6が小6で、図7が中3である。

図6も図7も、中央値を示す箱中央の横線が、各箱ひげ図で左から右にかけて上がっている。

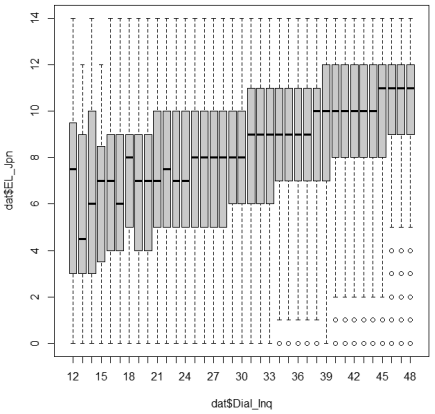
SES が低い児童生徒ほど、学力の得点が低い。同様に、学校で Dial_Inq をしていないと感じる児童生徒ほど、学力の得点が低い。

非認知「徳」（Virtus）と対話・探究学習（Dial_Inq）との関係を示す箱ひげ図を見ると、学校で対話・探究学習に恵まれる児童生徒とそうでない児童生徒との間の「非認知『徳』格差」は、深刻であることがわかる。学校で対話・探究学習に熱心に取り組めば取り組むほど、児童生徒の非認知「徳」スコアは「うなぎ上り」になることが、この分析から明らかである。

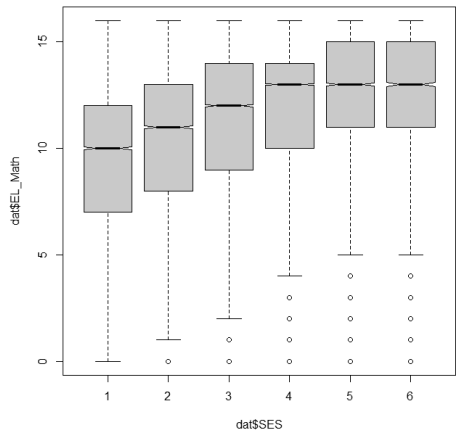
以上から結論として、対話・探究学習の学習効果は、学力に対して SES に匹敵するかやや大きめであり、非認知能力に対しては、SES とは比較にならないほど大きいと評価できる。SES には非認知能力に対する効果はない。



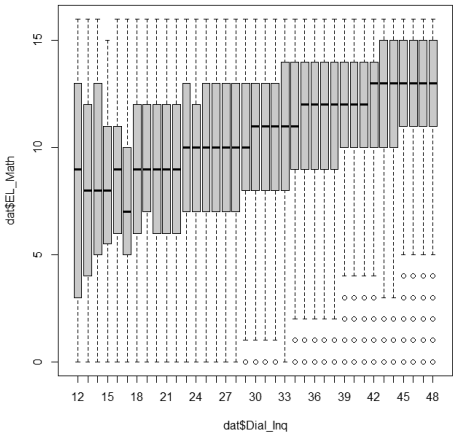
国語学力×SES



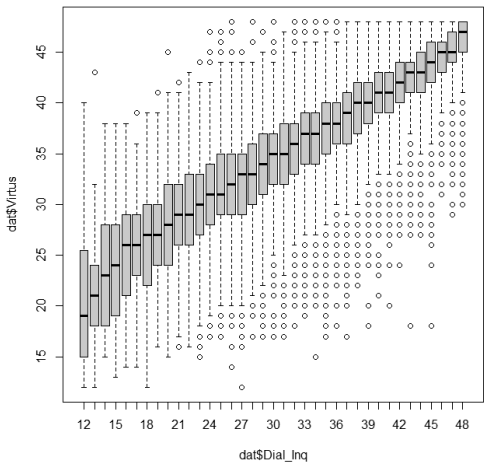
国語学力×対話・探究学習



算数学力×SES

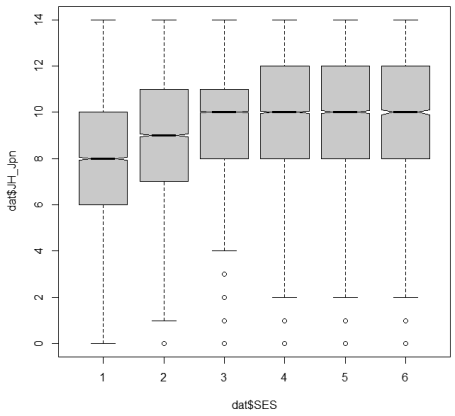


算数学力×対話・探究学習

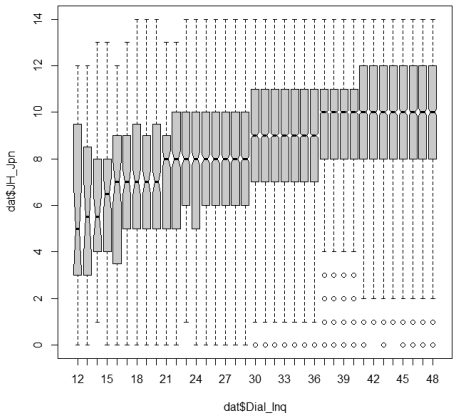


非認知「徳」×対話・探究学習

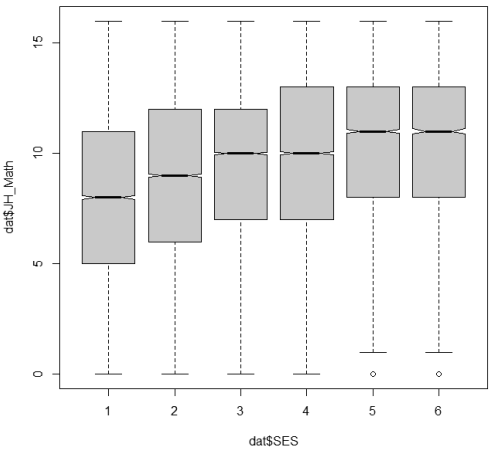
図６：令和３年度、小６の学力・非認知「徳」・SES・対話・探究学習の関係



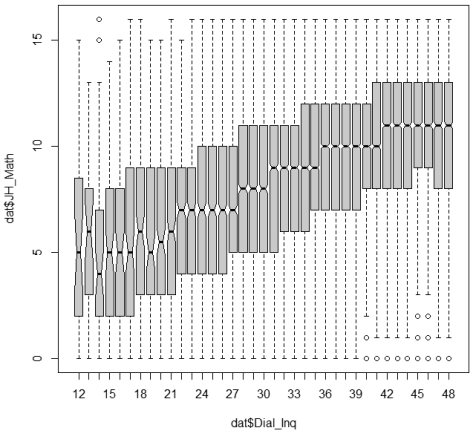
国語学力×SES



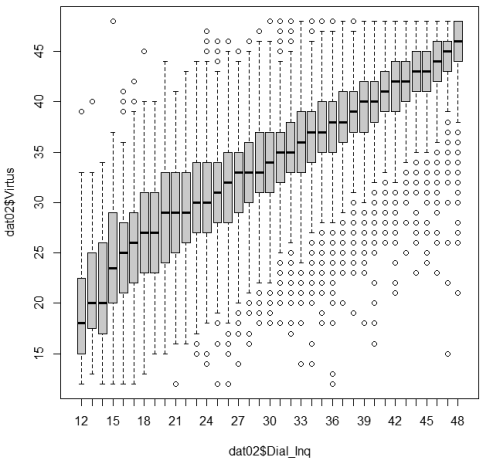
国語学力×対話・探究学習



数学学力×SES



数学学力×対話・探究学習



非認知「徳」×対話・探究学習

図７：令和３年度、中３学力・非認知「徳」・SES・対話・探究学習の関係

７. むすびに代えて

以上、本稿では、新しいシステムや合成変数を開発することで、対話・探究学習の一般的な価値を明らかにすることを試みた。「一般的な」とは、全国の小学６年生と中学３年生を母集団とした場合、という意味である。集団サイズを全国から、市町村規模とか学校規模に変えるならば、対話・探究学習の効果量も変わってくる。全国規模の効果量はあくまで目安である。基準値として利活用してほしい。

私たちが開発したものを改めて整理すれば、「平均値差検定システム」「全国学力・学習状況調査　平均ゾーンシステム」という２つのウェブシステムと、「非認知『徳』尺度」と「対話・探究学習尺度」の２つの合成尺度である。

ウェブシステム利用の際は、私たち DS-EFA のホームページの解説を参照いただきたい。効果量ｄの基準値としては、0.50を私たちは提案する。0.50未満の効果量であれば、実質的な平均差はないと評価する。

また今回の研究から、合成尺度の開発が、学習効果の可視化に有効であることがわかった。そこで私たちは、本研究の続編として、全国学力・学習状況調査結果を利活用し、非認知「徳」スコアの経年変化を各自治体や学校が把握できるようにするために、平成31年度、令和３年度、令和４年度に共通する質問項目に限定して、新たに「非認知『徳』尺度」「対話・探究学習尺度」を開発した。加えて「授業充実度尺度」という新しい合成尺度も開発した。この研究についても近く公表する。

さらに私たちは目下、教育委員会や学校に返却された調査結果のエクセルシートをコピー & ペーストすることで、これらの合成変数を計算し、全国と個別の自治体や学校とを比較するシステムを開発中である。

８. 補遺（Supplement）

本研究では、「3.「効果量」について」で述べたように、独自に効果量などを評価するウェブアプリを開発し一般に公開している。以下では、このウェブアプリについて補足しておく。教育現場において、統計の専門的な知識が十分ではない教員等も容易に統計の有意性や効果量を評価できるようにするために、特にニーズの高いと思われる平均値差の検定に特化した単機能のウェブアプリを作成し、「DS-EFA子ども教育データサイエンス」の HP 上に公開した。このウェブアプリは、前述の「平均ゾーンシステム」と同様に、javascript を用いて開発された。計算結果の数値を表示するとともに、必要なグラフなども出力するようにした。計算する統計値は、検定統計量ｔ値、確率Ｐ値、95%信頼区間、帰無仮説の棄却の可否である。さらに、効果量である Cohen の ｄ値と Hedges の ｇ値を計算できる点が大きな特徴である。平均値や標準偏差が既知であるケースは、「対応のないデータの平均値差の検定」と「対応のあるデータの平均値差の検定」として、また、元のデータを用いるケースは「ユーザーの元データを用いた平均値差の検定」として、独立なウェブアプリとして開発を行った。検定結果や効果量に加えて、与えられた平均と標準偏差に基づく正規分布のグラフを表示し、２つのグループの平均値の差を直感的に理解できるよう工夫した（図8）。ここでは、「平均ゾーンシステム」と同様に、JavaScript のグラフ作成ライブラリである Chart.js を利用した。

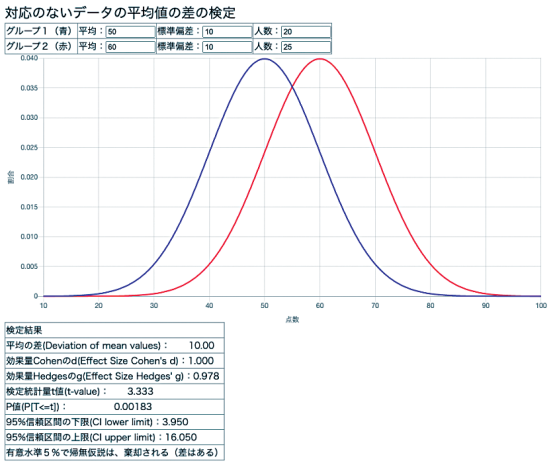


図8:「対応のないデータの平均値差の検定」のページ

「対応のあるデータ」の場合には、「対応のないデータ」と基本的な部分は同一であるが、それぞれのグループの点数分布を正規分布で表示するのに加えて、対応するデータの差の分布のグラフも同時に表示することにした（図9）。

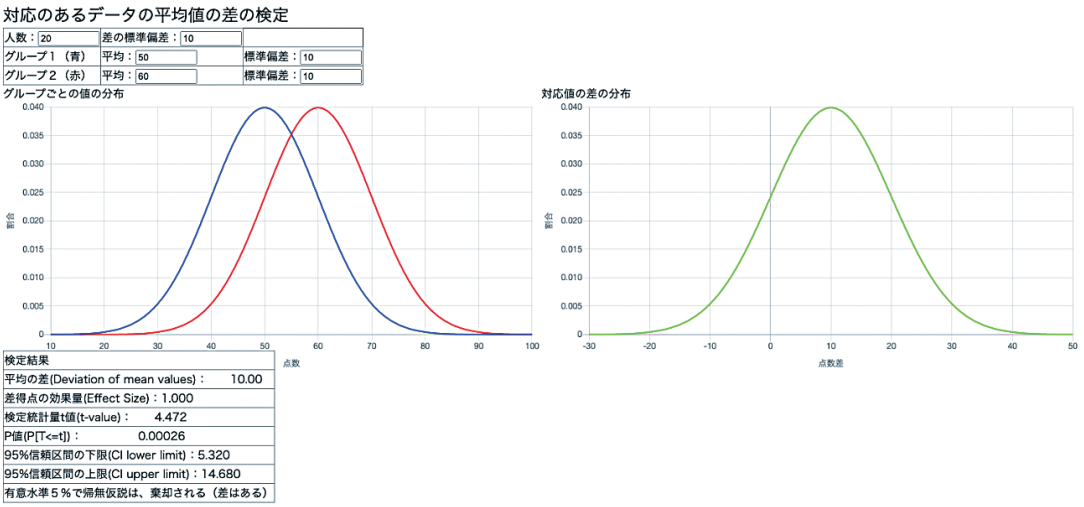


図9:「対応のあるデータの平均値差の検定」のページ

さらに、ユーザーが保有する元のデータから、直接的に平均値の差の検定を行うことができるウェブアプリも用意した。この方法では、ユーザーは2つのグループの元データをそれぞれの <textarea> 要素に、コピー / ペーストなどで入力するだけで良い。このとき、「対応するデータ」である場合には、データの並ぶ順番が対応していなければならない。ウェブアプリでは、この点の注意も与えている（図10）。

図 10:「ユーザーの元データを用いた平均値差の検定」のページ

「対応のないデータ」であるのか、「対応のあるデータ」であるのかは、自動的にデータ数をカウントし、ウェブアプリ上でその結果をメッセージで表示することとした。処理は両方のケースについて同時に行う仕組みにした。したがって、2つのグループのデータ数が同一であった場合には、「2つのデータ数は同じです。対応のないデータと対応のあるデータの両方で検定します」というメッセージを表示し、両方の検定結果が表示される。なお、ユーザーが入力するデータに空行などがあっても、それらを除外して正常に計算できるように工夫した。また、数値ではない文字列などが入力されている場合にも、エラーメッセージを表示するようにしている。このアプリでは、ユーザーの元のデータを扱っているので、本来の点数の分布が正規分布に近いのか否かは、データのヒストグラムを作成すれば確認できる。そこで、元の2つのグループのデータのヒストグラムと、「対応のあるデータ」のケースに対応した差のヒストグラムをそれぞれグラフとして作成し表示に加えた（図11）。このために、さらに別の JavaScript のグラフ作成ライブラリである plotly.js を使用した。plotly.js では、作成したグラフを拡大したり、ドラッグして動かしたり（パン）する様々な機能が付加されている。

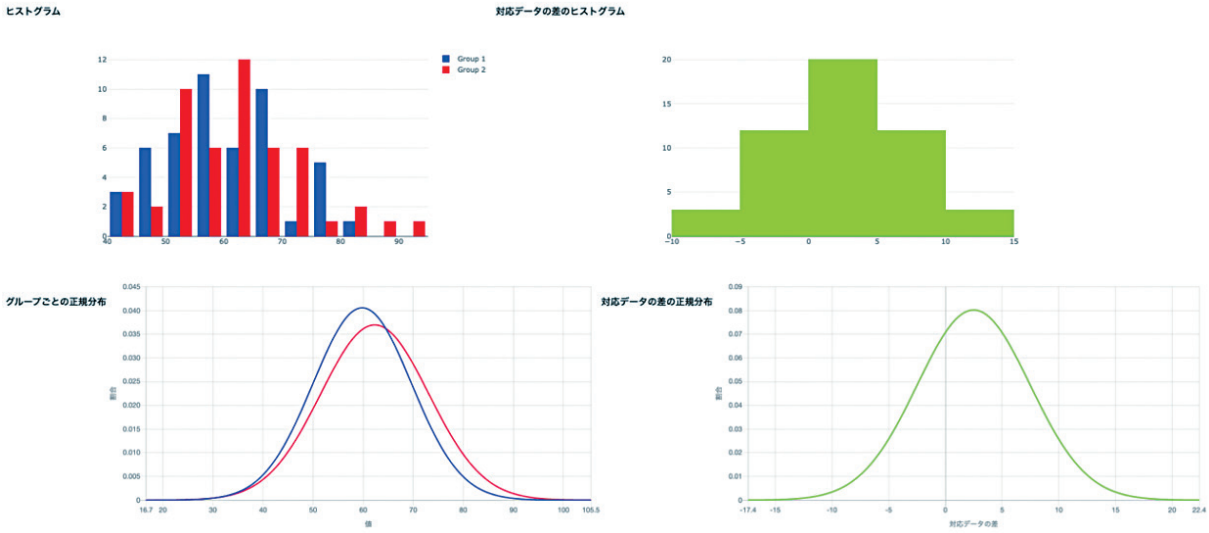


図11:「ユーザーの元データを用いた平均値差の検定」のページ後半部分（グラフ表示）

【付記１】

第１執筆者は本文を、第２執筆者は補遺(Supplement)を執筆した。第３執筆者はジョン・ハッティ名誉教授と連携し、ハッティ理論の本研究による理解の妥当性をチェックした。第４執筆者は「全国学力・学習状況調査『平均ゾーンシステム』」のウェブアプリを共同開発した。第５・第６・第７執筆者は「子どもの『徳』尺度」「対話・探究学習尺度」を共同開発し、第５執筆者が命名した。第６執筆者は教育心理学の専門的立場から、合成変数開発をスーパーバイズした。

【付記２】

本研究は科学研究費助成事業、基盤研究Ｂ「グローバル世界を視野とする学力・非認知能力の効果的学校モデル」（2020-22年度:20H01667）（研究代表者:田端健人）の研究成果の一部である。

【付記３】

本研究は令和３年度全国学力・学習状況調査の匿名データを文部科学省から貸与を受けた研究成果である。貸与の「ガイドライン」に従い、本発表内容について文部科学省に事前に報告し許可を得ている。
【付記４】本研究では、統計学やテスト理論の専門家である東北大学大学院教育研究科 柴山直教授から専門的技術指導をいただいた。この場をお借りして深く感謝申し上げる。

引用文献

Cohen, J.（1988）Statistical Power Analysis for the Behavioral Sciences, 2nd Edition. Routledge.
Cunha, F. & Heckman, J. J.（2006）Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation, The Journal of Human Resources, XLIII（4）: 738-782.
DS-EFA（2022）平均値差検定システム . https://ds-efa.info/cohensd/ ［2022.09.28最終閲覧］
ハッティ , J.（2018）学習に何が最も効果的か―メタ分析による学習の可視化:教師編―. 原田信之訳者代表 . あいり出版 .
Heckman, J. J. & Rubinstein, Y.（2001）The Importance of Noncognitive Skills: Lessons from the GED Testing Program. The American Economic Review, 91（2）: 145-149.
Heckman, J. J., Stixrud, J.& Urzua, S.（2006）The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior, Journal of Labor Economics 24（3）: 411-482.
Heckman, J., Pinto, R. & Savelyev, P.（2013）Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes, American Economic Review 103（6）: 2052-2086.
川口俊明（2006）学力格差と「学校の効果」―小学校の学力テスト分析から―. 教育学研究, 73-4:350-362.
経済協力開発機構（OECD）（編著）（2019）社会情動的スキル―学びに向かう力―. 無藤隆・秋田喜代美（監訳）. 明石書店 .
小林雄一郎・濱田彰・水本篤（2020）R による教育データ分析入門 . オーム社 .
新潟県教育委員会（2004）平成16年度「全県学力調査」報告書 .
新潟県教育委員会（2006）平成18年度「全県学力調査」報告書 .
村井潤一郎・橋本貴充（2018）統計的仮説検定を用いる心理学研究におけるサンプルサイズ設計 . 心理学評論, 61-1:116-136.
大久保街亜・岡田謙介（2012）伝えるための心理統計―効果量・信頼区間・検定力―. 勁草書房 .
柴山直（2006）古典的テスト理論 . 大沢武志ほか（編）人事アセスメントハンドブック . 金子書房, pp.453-481.

志水宏吉(2006)学力格差を克服する学校―日本版エフェクティブ・スクールを求めて―.教育学研究, 73-4:336-349.
田端健人(2021)子どもの言葉データサイエンス入門―形態素解析システム jReadability の活用と検証―.パイデイア出版.
田端健人・丸山千佳子・本図愛実・原田信之・野坂実央(2022)第2版 全国学力・学習状況調査を有効活用する「平均ゾーンシステム」の開発―Z検定と効果量の可視化―.田端健人(編著) IRT分析ソフト EasyEstimation による全国学力・学習状況調査の検証と経年比較.パイデイア出版, pp.115-135.
Wasserstein, R. L. & Lazar, N. A.(2016)統計的有意性とP値に関する APA声明.佐藤俊哉訳.
https://www.biometrics.gr.jp/news/all/ASA.pdf
[2022.09.28最終閲覧]