参考資料3

令和５年度委託事業
学力調査を活用した専門的な課題分析に関する調査研究
（専門的な知見を活用した高度な分析に関する調査研究）報告書

【テーマ】
A. 令和４年度全国学力・学習状況調査の理科の結果を活用した専門的な分析
・我が国の児童生徒の理科の学力や学習状況に関する傾向等の分析

AFORCE

国立大学法人
宮城教育大学
Miyagi University of Education

内容

事業名：令和５年度「学力調査を活用した専門的な課題分析に関する調査研究(専門的な知見を活用した高度な 分析に関する調査研究)」

調査研究テーマ

A.令和４年度全国学力・学習状況調査の理科の結果を活用した専門的な分析

・我が国の児童生徒の理科の学力や学習状況に関する傾向等の分析

調査研究の内容

令和４年度全国学力・学習状況調査の理科の教科調査や質問紙調査の結果を用いて、我が国の児童生徒の理科に関する傾向等の詳細な分析を行う。その際、国際学力調査(TIMSS・PISA 等)における我が国の結果と比較した分析や、小学校及び中学校の理科における男女差の分析を行うとともに、男女差が見られる場合にはその要因の分析を行う。

はじめに

令和４年度全国学力・学習状況調査の教科に関する調査（教科調査）において、理科の調査が実施された。理科の調査は平成 24 年度から実施されており、平成 27 年度、平成 30 年度に続いて４回目であった。

本調査研究においては、過去４回の全国学力・学習状況調査の理科の教科調査や質問紙調査の結果を用いて、TIMSS や PISA といった国際的な学力調査と我が国の児童生徒の理科に関する傾向等の分析を行った。特に、日本において、昨今、理系に進学する女子が少ないという話題への関心が高まっていることから、理科の教科調査や質問紙調査の結果に見られる男女差に着目した。

主な分析の観点は以下のとおりである。

１．理科の平均値の経年変化分析

全国学力・学習状況調査における理科の学力値に、実質的な向上や低下は認められるか、平成 24、27、30、令和４年度の小学６年と中学３年の経年変化を分析した。また、この経年変化に係る分析結果が、TIMSS や PISA といった国際的な学力調査における経年変化の状況とどれほど整合的かについても分析した。

２．理科の教科調査の男女比較

全国学力・学習状況調査の教科に関する調査で、理科の男女の平均値に実質的な差はあるかについて分析するため、平成 24、27、30、令和４年度の小６と中３の男女の平均値差を検証した。また、その男女比較の結果が、国際的な学力調査の男女比較の結果と整合的かどうかについても分析した。

３．教科間の相関の男女比較

全国学力・学習状況調査の理科、国語、算数・数学の教科間において、正答率の相関の強さに男女差はあるのかについて分析した。

４．理科に関する質問紙調査の回答傾向の男女比較

理科の学習に対する興味・関心や授業の理解度等の男女比較を行った。また、国語、算数・数学についても同様の男女比較を行い、理科の結果と対比した。

５．その他

平成 24、27、30、令和 4 年度の全国学力・学習状況調査理科調査のテスト精度についても検討した。

本委託研究は、株式会社エーフォースが受託し、全国学力・学習状況調査の理科に焦点化した分析を宮城教育大学チームに依頼した。その分析結果を基に、同大学大学院教育学研究科の田端健人が本報告書を執筆した。

体制

企画、調整、取りまとめ

大塩喬介　　株式会社エーフォース　営業部長

吉田裕之　　株式会社エーフォース

分析、報告書執筆

田端健人　　国立大学法人宮城教育大学大学院教育学研究科　教授

分析協力

市瀬智紀　　国立大学法人宮城教育大学教育学部　教授

平真木夫　　国立大学法人宮城教育大学大学院教育学研究科　教授

本図愛実　　国立大学法人宮城教育大学大学院教育学研究科　教授

板垣翔大　　国立大学法人宮城教育大学教育学部　准教授

山田美都雄　国立大学法人宮城教育大学アドミッションオフィス　准教授

１．理科の平均値の経年変化分析

―国内および国際学力調査（TIMSS、PISA）における検討―

１．１．IRT 分析ソフトによる経年比較―国内調査での検討―

リサーチクエスチョン：

全国学力・学習状況調査の教科に関する調査の経年で、理科の学力値に実質的な向上や低下は認められるか？

分析結果：

あくまで実用的レベルでの経年変化分析であるが、平成 24、27、30、令和 4 年度の小学 6 年生および中学 3 年生の理科学力値に、実質的な変化は認められない。

分析方法：

平成 24、27、30、令和 4 年度の小学 6 年生（以下「小 6」）と中学 3 年生（以下「中 3」）の経年変化を分析する。

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、受検集団と難易度や識別力の異なるテスト結果を単純に比較することはできない。共通の問題もないため、原理的には尺度等化ができず、経年変化を見ることができない。学術的に厳密な経年比較をするためには、文部科学省が令和 4 年度、東北大学（代表：柴山直教授）に委託して実施した「経年変化分析調査」のような特別な調査分析が必要である。

この原則を抑えたうえで、実用的なレベルでの経年比較分析として、東北大学の熊谷龍一氏が開発した IRT 分析ソフト EasyEstimation[1]を使った分析方法[2]を、本調査研究で活用する。

この分析方法は、次の仮定と推定をもとになされている。

---

[1] Cf., 熊谷龍一(2009), 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, 107-118. 熊谷龍一(2012), 統合的 DIF 検出方法の提案－“EasyDIF”の開発－. 心理学研究, 83, 35-43. 本ソフトは、以下の URL からダウンロードできる。

https://irtanalysis.main.jp/

なお EasyEstimation は、令和 3 年度「全国学力・学習状況調査」経年変化分析調査において、IRT 分析に利用されている（cf., 文部科学省(2022a), 令和 3 年度「全国学力・学習状況調査」経年変化分析調査テクニカルレポート, 付録 A, B)。

令和3年度『全国学力・学習状況調査』経年変化分析調査テクニカルレポート：

https://www.nier.go.jp/21chousakekkahoukoku/kannren_chousa/pdf/21keinen_tech_01.pdf

令和3年度『全国学力・学習状況調査』経年変化分析調査テクニカルレポート別冊 （標本抽出方法）：

oku/kannren_chousa/pdf/21keinen_tech_02.pdf

[2] Cf., 田端健人編著(2022), 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.

1. 複数年度が同じ標準正規分布であると仮定する。母集団分布に関する初期値を複数年度同じもの（＝標準正規分布）にそろえる。
2. その初期値のもとで、それぞれの年度の母集団分布（平均や標準偏差）を推定する。
3. 得られた推定値を複数年度で比較する。

初期値が同じため、複数年度の推定値が大きく変化することはない、という限界がある。しかし、複数年度の実際の共通情報がないため、仮想的に初期値をそろえることで等化の代替をしていることになる。原理的限界があるものの、実際の分析結果は「経年変化分析調査」の結果ともほぼ整合的である[3]。

利用データ：

IRT 分析ソフトで母集団母数の推定を行うため、悉皆調査の個票データ約 100 万件[4]の 10%程度の無作為サンプリングデータで十分と考えられるが、今回は個票データの全数を利用した。

欠測値（空白）を削除した後のデータ処理数は、表 1-1 の一覧の通り。

年度は元号の頭文字のアルファベット、小 6 は「EL」、中 3 は「JH」の略記号を用いた。

表 1-1：EasyEstimation での処理件数

	H24	H27	H30	R4
EL	262,899	1,081,705	1,049,832	1,006,852
JH	443,243	1,062,762	1,011,878	941,669

単位：人

分析結果の可視化：

[3] Cf., 田端(2022), pp.23-29.

[4] 平成 24 年度調査は「悉皆」ではなく、「抽出」調査であった。そのため、表 1-1 の通り、処理件数も少なくなっている。

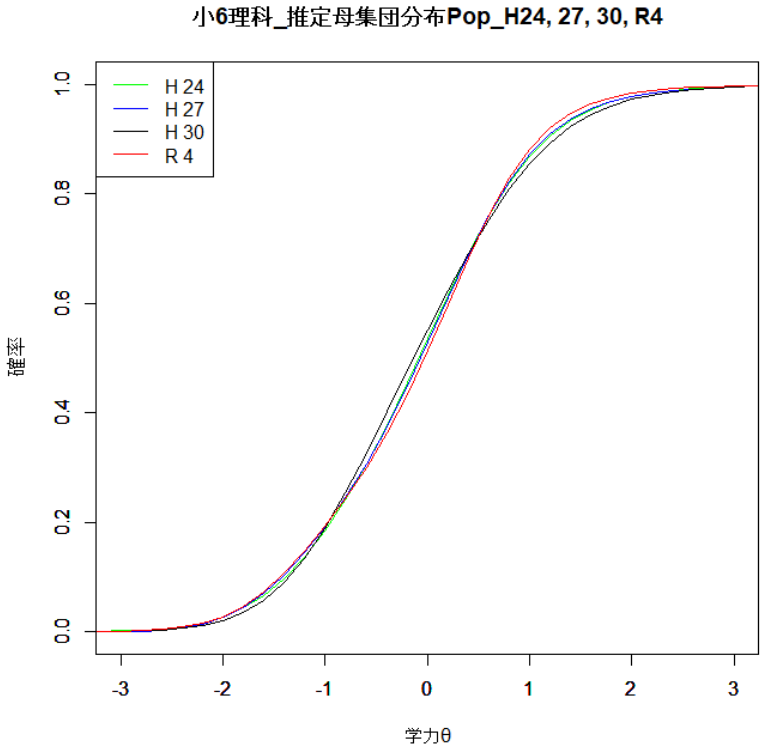


図 1-1：小 6 理科「累積分布」の経年比較

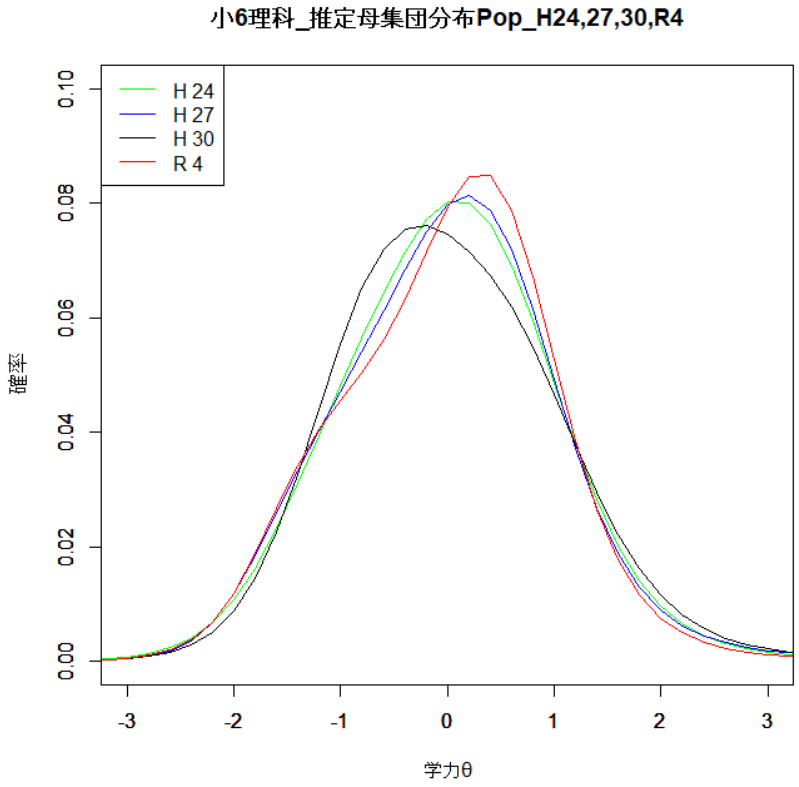


図 1-2：小 6 理科「度数分布」の経年比較

累積分布（図 1-1）でみると、平成 24、27、30、令和 4 年度の 4 か年度の差はほとんど見えない。

度数分布（図 1-2）でみると、これら 4 か年度に若干の差が見える。

そこで、度数分布曲線の形状が最も異なる平成 30 年度（黒）と令和 4 年度（赤）の平均値差を、効果量（Cohen の d 値）[5]で見積もる。基準値は d≧0.40 を「小さいが差がある」とする [6]。EasyEstimation の母集団分布推定結果で、平成 30 年度は平均 0.016、標準偏差 1.033、令和 4 年度は平均 0.004、標準偏差 0.995 であり、差分効果量 d=0.012 となり、差はないに等しい。サンプル数を 1 万件と仮定して t 検定をしても有意水準 5%で帰無仮説 [7]が棄却される（有意差なし）。

【結論】あくまで実用的レベルの経年比較だが、平成 24、27、30、令和 4 年度の間に、わが国小学 6 年生の理科の学力値に、実質的な向上や低下は認められない。

同じ要領で、中 3 理科の経年変化を分析する。分析結果のグラフは図 1-3 と図 1-4 になる。

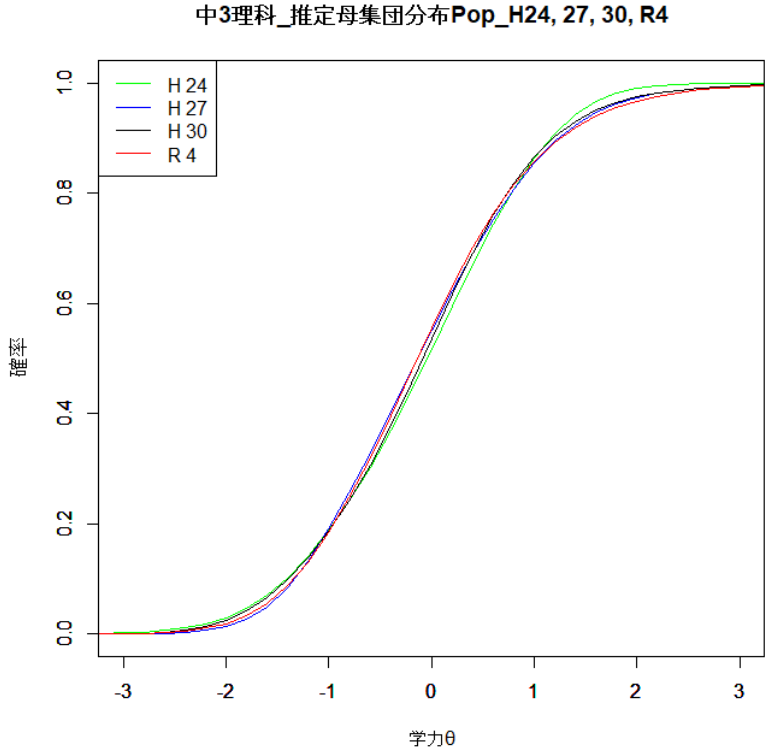


図 1-3：中 3 理科「累積分布」の経年比較

[5] 「効果量」については、本章末の用語解説 A を参照されたい。

[6] Cf.，田端健人(2023)，「教育の現象学」のデータサイエンス的転回―全国学力・学習状況調査結果の分析から―．学ぶと教えるの現象学研究，20，pp.97-100．またこの基準値については、本章末の用語解説 A を参照されたい。

[7] 「帰無仮説検定」については、本章末の用語解説 A を参照されたい。

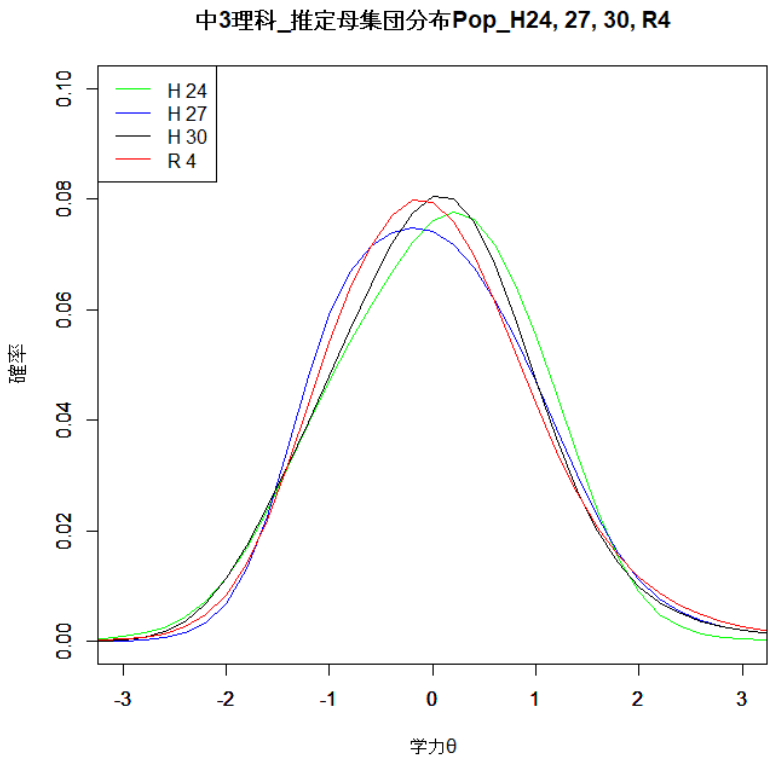


図 1-4：中 3 理科「度数分布」の経年比較

累積分布（図 1-3）でみると、平成 24、27、30、令和 4 年度の 4 か年度の差はほとんど見えない。

度数分布（図 1-4）でみると、これら 4 か年度の差が若干見えるが、小 6 の度数分布グラフ（図 1-2）と比較すれば、実質的な差とは評価できない。

試みに、平均差が最も大きい平成 27 年（紫）と平成 24 年（緑）の効果量を計算すると、EasyEstimation の母集団分布推定結果で、平成 27 年の平均は 0.024、標準偏差は 1.016、平成 24 年の平均は 0.012、標準偏差は 0.993 であり、差分効果量は d=0.012 となり、小 6 の場合と同じく差はないに等しい。

【結論】あくまで実用的レベルの経年比較だが、平成 24、27、30、令和 4 年度の間に、わが国中学 3 年生の理科の学力値に、実質的な向上や低下は認められない。

１．２．国際調査（TIMSS、PISA）や国内他調査との整合性の検討

リサーチクエスチョン：

上記１．１の結論は、理科の学力の経年変化に関する国際調査とどれほど整合的か。TIMSS や PISA の長期トレンド評価や、文部科学省による経年変化分析調査の結果と、それは整合的か否か。

結論：

本調査研究の１．１．の結論「わが国の小 6 と中 3 の理科の平均値の長期トレンドに、実質的な変化はない」は、PISA の科学的リテラシーの OECD による長期トレンド評価と整合的であり、かつ TIMSS の小 4 と中 2 の理科調査の長期トレンドに関する本調査研究の独自評価とも整合的である。

また、本調査研究１．１．の結論は、文部科学省による令和3年度経年変化分析調査の結果と両立可能であり、これにより反証される結論ではない。

大規模国際学力調査TIMSS（Trends in International Mathematics and Science Study: IEA国際数学・理科教育動向調査）やPISA（Programme for International Student Assesment：OECD生徒の学習到達度調査）も理科の学力に関し、日本の児童生徒の経年変化を分析・報告している。

例えば、TIMSS2019の報告書は、小学校4年生では、2019年の平均得点は2015年に比べ「統計的に有意に低い」と評価し、2007年や2003年に比べると「統計的に有意に高い」と評価している[8]。また中学2年生では、2019年の平均得点は、2011年や2007年や2003年に比べ「統計的に有意に高い」と評価している[9]。

一方、15歳（日本の場合は高校1年生に相当）を対象としたPISA2018の報告書では、理科にあたる科学的リテラシーの平均得点の経年変化として、日本の場合、2018年は2015年や2012年に比べ低く、その差には「統計的な有意差がある」と評価している[10]。ところが、「OECD 生徒の学習到達度調査2018年調査（PISA2018）のポイント」では、2000年から2018年のおよそ20年間の長期トレンドとして、日本は「平均得点のトレンドに統計的に有意な変化がない」と評価している[11]。「OECD生徒の学習到達度調査―PISA2022のポイント―」でも、「日本とOECDの平均得点の推移（調査開始時-2022年）」で、「OECDは平均点の長期トレンドが下降しているが、日本は平坦型（平均得点のトレンドに統計的に有意な変化がない）」としている[12]。

このように、同じOECDの評価でも2か年比較と長期トレンド評価とに不整合がみられる。これはどのような統計的処理により有意差を求めるかによるのであろう。また調査対象年齢や調査問題に違いがあるとはいえ、TIMSSとPISAの2か年比較の評価にも不整合がみられる。理科の平均得点の推移だけを見ても、図1-5から視覚的に明らかなように、2011-12年を起点にすれば、2015-18にかけて、TIMSSは小4と中2とも上昇傾向だが、逆にPISAは下降傾向である。

[8] Cf., 国立教育政策研究所編(2021), TIMSS2019 算数・数学教育／理科教育の国際比較―国際数学・理科教育動向調査の2019年調査報告書, 明石書店, p.172.

[9] Cf., 国立教育政策研究所編(2021), p.173.

[10] Cf., 国立教育政策研究所編(2019), 生きるための知識と技能―OECD生徒の学習到達度調査（PIISA）2018年調査国際結果報告書―, 明石書店, p.193.

[11] https://www.nier.go.jp/kokusai/pisa/pdf/2018/01_point.pdf ［2024/1/22最終閲覧］

[12] https://www.nier.go.jp/kokusai/pisa/pdf/2022/01_point_2.pdf ［2024/1/22最終閲覧］ 括弧内原文。

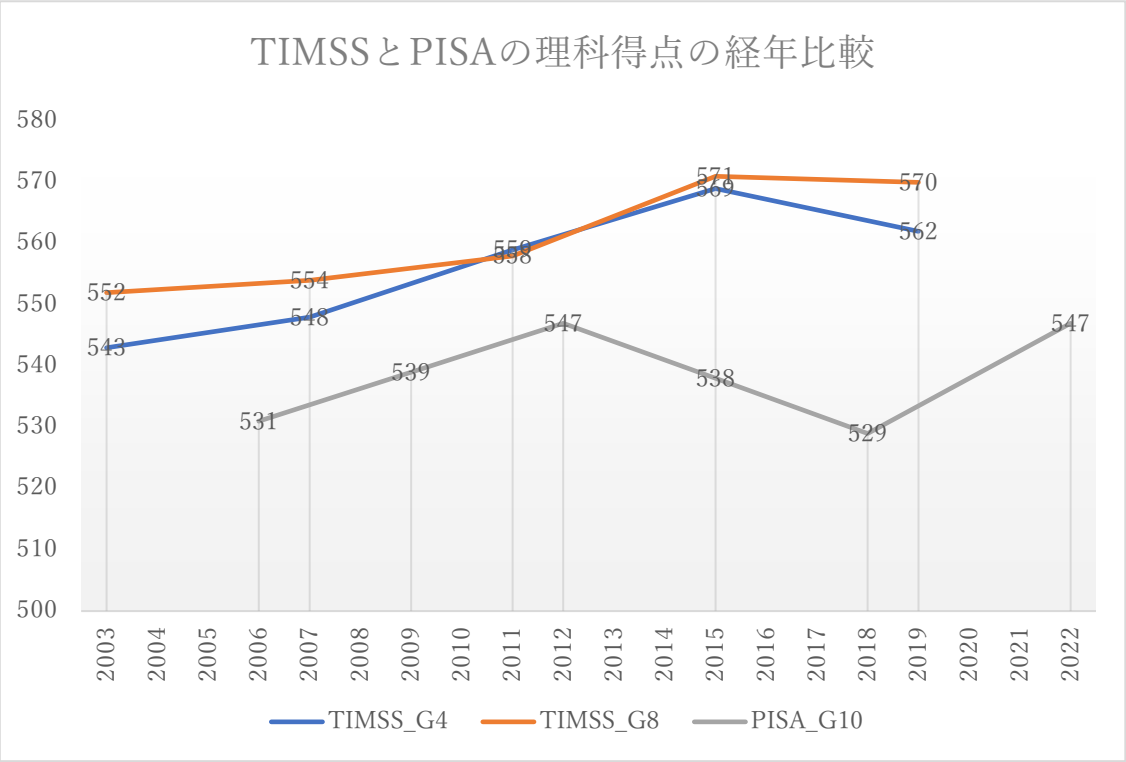


図 1-5：TIMSS と PISA の理科平均得点の経年比較（公表データから筆者作成）

本調査研究１．１．の結果は、PISA の長期トレンド評価と整合的である。

TIMSS も長期トレンド評価をするならば、整合的な結果が得られるだろうか。TIMSS は各国・地域の長期トレンド評価を公表していないため、本調査で独自に評価を試みる。

【検証 01】　TIMSS の理科平均得点の長期トレンドの評価方法：

1. 本調査研究１．１と同じ効果量 d 値による評価。
2. d＜0.40 なら「実質的な差はない」と評価。
3. TIMSS2003 から TIMSS2019 までの平均得点の最小値と最大値の差の効果量を計算。図 1-5 より、平均値差が最も大きいのは、G4（小 4）で、2015 年が最大値 569、2003 年が最小値 543、平均値差は 26 ポイントである。
4. 効果量 d の計算には平均値と標準偏差が必要だが、TIMSS の報告書では標準偏差が記載されていない。そこで、TIMSS2003 と TIMSS2015 の小 4 の平均値と標準偏差を、本調査研究独自に算出。
5. 分析ソフトは、TIMSS の実施主体である IEA（International Association for the Evaluation of Educational Achievement：国際教育到達度評価学会）が公開している IDB アナライザーV5 を使用。
6. PV’s（Plausible Values：推算値）と生徒ウェイトを使い、母集団推計を行う。
7. その結果から「平均値差検定システム（AVES）」にて効果量を計算。

分析結果

表 1-2：TIMSS 小学 4 年生、理科調査平均得点の最大値と最小値の差分効果量

	サンプル数	推定母集団	平均値	標準偏差	TIMSS2015-TIMSS2003	平均値差の効果量 d
TIMSS2015_G4_Sci	4,383	1,061,010	569.01	64.95	25.54	0.37
TIMSS2003_G4_Sci	4,535	1,172,766	543.47	73.12		

平均値差の効果量 d=0.37 であるため、実質的な差はないと評価。

ただし、効果量 0.37 はそれなりの差ではある。そこで、OECD の長期トレンド評価で「統計的に有意な変化はない」とした PISA の科学的リテラシーのわが国の平均得点の差が大きい 2022 年（547 ポイント）と 2018 年（529 ポイント）の効果量も計算する。

表 1-3：PISA 科学的リテラシー平均得点の最大値と最小値の差分効果量

	サンプル数	推定母集団	平均値	標準偏差	PISA2022-PISA2018	平均値差の効果量 d
PISA2022_Sci	5,760	1,021,370	546.63	92.97	17.49	0.19
PISA2018_SCI	6,109	1,078,921	529.14	92.07		

科学的リテラシーでは、最も平均得点差が大きな 2022 年と 2018 年も、ポイント差で 17.49、平均値差の効果量で d=0.19 である。OEDC の長期トレンド評価では、平均差の効果量にして 0.19 は「統計的に有意な変化ではない」となる。PISA で、平均得点差がいっそう大きな教科や年度はないだろうか。

平均得点差が最も大きいのは、読解力で、2003 年および 2006 年の平均得点は 498、2012 年は 538 で、その差は 40 ポイントである。これを効果量で見積もると、どうなるか。独自に標準偏差を算出し、効果量を計算する。

表 1-4：PISA 読解力平均得点の最大値と最小値の差分効果量

	サンプル数	推定母集団	平均値	標準偏差	PISA2022-PISA2018	平均値差の効果量 d
PISA2012_Read	6,351	1,128,179	538.05	98.69	40.09	0.399
PISA2006_Read	5,952	1,113,701	497.96	102.39		

効果量は、小数第 3 位を切り捨てれば 0.39 となり、四捨五入すれば 0.40 になる。微妙な値であるが、OECD の長期トレンド評価からすると、効果量で 0.40 程度は実質的な変化とはみなされていない。

以上から、TIMSS 理科の 2015 年と 2003 年の平均得点差の効果量 0.37 は、長期トレンドとして「実質的な変化がある」という根拠にはならない。逆に、実質的な平均値差効果量を 0.40 以上とする本調査研究の評価基準は、それなりの妥当性をもつことになる。

【検証 02】　文部科学省による経年変化分析調査との整合性

1.　「令和 3 年度『全国学力・学習状況調査』経年変化分析調査テクニカルレポート」（文部科学省）

[13]では、「平成 28 年度分布を基準に令和 3 年度の変化の有無をみる」(p.67) とし、小 6 と中 3 の国語ならびに算数・数学の経年変化を統計学的に厳密に検証している。

2. 検証の結果、同レポートは、「小学校・中学校ともに国語に関しては、学力スコア分布（累積相対度数分布）の状況は両年度でほとんど変化は観察されなかった」とし、「国全体としてみれば、児童生徒の学力の低下や向上といった変化は認められなかった」と評価している（p.67)。
3. 一方、「算数・数学については、令和 3 年度の学力スコア分布（累積相対度数分布）は基準である平成 28 年度の学力スコア分布の右側に（全体的にみて学力スコアが高い方へ）若干移動していることが観察できる」(p.67) としている。ただその評価については、「国全体でみれば、算数・数学について若干学力が向上しているとも解釈しうるが、・・・(中略)・・・全国の学力分布の状況の変化の有無は中長期的に継続分析する必要があることを踏まえ、次回（令和 6 年度予定）以降の結果もあわせて引き続き分析していくこととする」(p.67) と述べ、学力の若干の向上の解釈の余地を残しながら、断言を控えている。
4. 理科や英語については、経年変化分析調査のための重複テスト分冊法による抽出調査がまだ整っておらず、全国の学力分布の変化の有無は検証されていない。
5. 以上のように経年変化分析調査は、平成 28 年度を基準として令和 3 年度の国語について、小学校・中学校ともに変化なしとし、算数・数学については、「若干」の向上が認められるものの明らかな向上との断言を控える程度である。教科の違いを考慮しても、小学校・中学校の理科で明らかな向上があると推測する根拠は乏しい。全国学力・学習状況調査の先行研究[14]において、異教科間の学力の相関係数がそれなりに高いとの知見がある。また本調査分析でも、国数理の相関係数は、大半が 0.7 を上回っている（第 3 章参照)。これらからすると、理科でのみ目立った経年変化があるとは推測し難い。
6. それゆえ、本調査研究１．１．の結論は、少なくとも、令和 3 年度経年変化調査分析の結果と相いれないものではない。令和 3 年度経年変化調査分析の結果は、合理的に考えて、本調査研究１．１．を必ずしも反証するものではない。

[13] https://www.nier.go.jp/21chousakekkahoukoku/kannren_chousa/pdf/21keinen_tech_01.pdf [2024/1/22 最終閲覧]

[14] 問題数が多かった平成 28、30 年度の異教科間の相関係数の平均は r=0.74、問題数が減少した令和 3，4 年度の異教科間の相関係数の平均は r=0.70 である（cf.，田端健人(2024)，全国学力・学習状況調査「教科に関する調査」の品質検証―平成 28、30 年度、令和 3,4 年度の比較―，宮城教育大学紀要，58（近刊））．

○　用語解説 A

A-1 効果量

「効果量（effect size)」とは、効果の大きさを表す指標のことである。標準化されており、測定単位に依存しないため、「異なる測定方法を用いた異なる研究から得られたデータを比較」[15]できる。それゆえ、特定の教育的働きかけの効果に関する複数の数量的研究の結果を統合して、その働きかけの全体としての効果量を算出する**メタ分析**でも用いられる。教育の統合メタ分析で近年注目を集めるジョン・ハッティの一連の研究[16]も、効果量を土台にしている。

効果量は大きく 2 種類に区別できる。1 つは、「2 つの群の間にどの程度の違いがあるか」を表わす「d 族（d family）の効果量」である[17]。本調査研究で利用する「コーエンの d 値（Cohen's d)」がこれに当たる。もう 1 つは、「2 つの変数間の関係がどの程度大きいか（もしくは小さいか)」を表わす「r 族（r family）の効果量」である[18]。よく知られているピアソンの積率相関係数 r がそれに当たる。効果量 d と相関係数 r との間には、次の関係式が成り立つとされる[19]。

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

効果量 d は、2 群の平均値と標準偏差とサンプル数がわかれば、次の式により算出できる[20]。グループ 1 の平均値をM_1、標準偏差をS_1、サンプル数n_1、グループ 2 の平均値をM_2、標準偏差をS_2、サンプル数n_2とする。

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

分母のS_pは、「プールした標準偏差」のことで、

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

で定義される。

効果量の強みは、サンプル数の影響をほとんど受けない点にある。効果の有無の検定で従来多用されてきた帰無仮説検定はサンプル数の影響を受け、サンプル数が大きくなるほど p 値が小さくなり、5%水準で有意差ありになる。例えば、偏差値で 1.0 の差の 2 群（偏差値 50 のグループと偏差値 51 のグループ）を仮定すると、両群のグループ人数が 769 名なら有意水準 5%で有意差なしだが、770 名以

---

[15] 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，p. 90. Cf.，大久保街亜・岡田謙介(2022)，伝えるための心理統計―効果量・信頼区間・検定力―，勁草書房，p. 47.

[16] Cf.，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. Cf.，Hattie，J.（2023)，*Visible Learning: The Sequel,* Routledge.

[17] 大久保・岡田(2022)，p. 46.

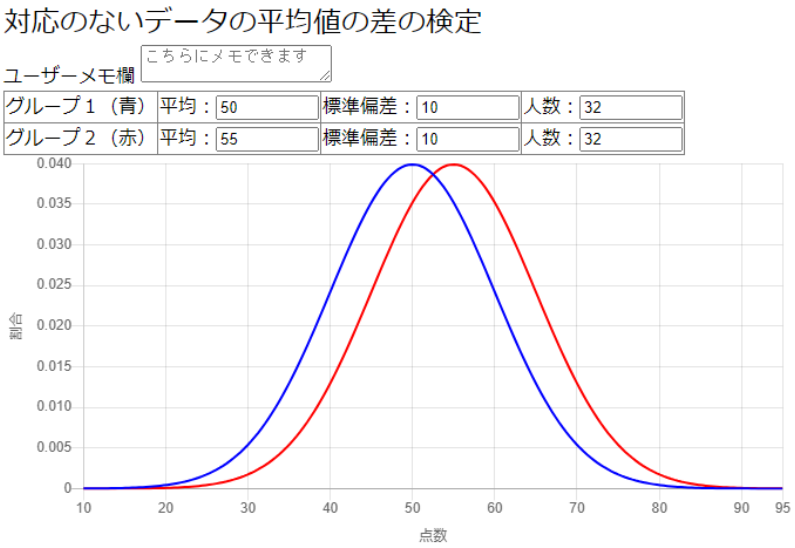
[18] 大久保・岡田(2022)，p. 46. 括弧内原文。

[19] Cohen，J.（1988)，*Statistical Power Analysis for the Behavioral Sciences, 2nd Edition*, Routledge，p. 23.

[20] Cf.，大久保・岡田(2022)，p. 55.

上になると有意差ありとなる[21]。両群の平均値差は、効果量で見積もれば d=0.10 であり、この値はサンプル数に依存しない。先の「プールした標準偏差」の式に 2 群のサンプル数（n_1 と n_2）が組み込まれているが、影響するのは 2 群のサンプル数の比であり、数量ではない。

なお、効果量 d は偏差値に換算でき、d=0.1 は偏差値 1.0、d=0.5 は偏差値 5.0、d=1.0 は偏差値 10 の差に対応する。図 1-6 は、偏差値 50 と偏差値 55 の 2 群（平均値差の効果量にして d=0.5）の分布曲線である。DS-EFA 開発の平均値差検定システム（AVES）[22]の描画と検定結果である。t 検定をすれば、2 群の人数が 31 名以上で 5%水準で有意差ありになる。



検定結果
平均の差(Deviation of mean values)： 5.00
効果量Cohenのd(Effect Size Cohen's d)：0.500
効果量Hedgesのg(Effect Size Hedges' g)：0.492
検定統計量t値(t-value)： 2.000
P値(P[T<=t])： 0.04989
95%信頼区間の下限(CI lower limit)：0.003
95%信頼区間の上限(CI upper limit)：9.997
有意水準５%で帰無仮説は、棄却される（差はある）

図 1-6：偏差値 50 と偏差値 55 の 2 群の分布曲線と検定結果

A-2 基準値

2 群の平均値差を効果量で測定する場合、効果量がどの程度であれば、「差がある」と言えるのか。あるいは効果量がどの程度であれば、その差は「小さい」とか「中くらい」とか「大きい」と評価できるだろうか。これは「基準値（reference value)」の問題である。基準値は、客観的絶対的に定めることができない。主観的判断が不可避である。

---

[21] 宮城教育大学 DS-EFA チーム開発の平均値差検定システム（AVES）でのシミュレーション結果より。シミュレーションの仕方は、まずグループ 1 に平均 50、標準偏差 10、人数 769 を、グループ 2 に平均 51、標準偏差 10、人数 769 を入力する。すると「有意水準 5%で帰無仮説は、棄却されない（差はない)」の結果となる。しかし両群の人数を 770 に増やすと、「棄却される（差はある)」の結果になる。

[22] https://ds-efa.info/script/cohensd.html

帰無仮説検定において、広く基準値とみなされる「有意水準 5%（p=0.05）」も、「あくまでも慣習であり、研究分野や領域によって変化」[23]する。効果量の基準値にも諸説ある。コーエンは、d=0.2 を「小さな差」、d=0.5 を「中くらいの差」、d=0.8 を「大きな差」とする目安を提案している[24]。しかし、d=0.2 程度の効果量でも、「現実に意味があると考えられる差の例は多数」[25]ある。コーエンがあげる例では、15 歳と 16 歳の女子の身長差がそれに当たる[26]。対して、d=0.5 の効果量は、14 歳と 18 歳の女子の身長差に当たる[27]。ハッティは、独自の統合メタ分析から d=0.4 を基準値とし、これは学校で 1 年間に達成できる平均的な値としている[28]。彼はまた続編で、教育に関する統合メタ分析の平均値として d=0.42 を基準値としている[29]。

本調査研究のリサーチクエスチョンは、「全国学力・学習状況調査の教科に関する調査で、平成 24 年度から令和 4 年度のおよそ 10 年の間に、わが国の小学 6 年生や中学 3 年生の学力値に、学力向上とか学力低下と言えるほどの、実質的な平均値差はあるのか」である。本調査研究を実施した宮城教育大学分析チームは、これまでの検討結果から[30]、d=0.4 を、実質的な差があると評価する最低ラインとした。さらに本調査研究では、2003 年から 2022 年までの TIMSS と PISA の日本の児童生徒の平均スコアの最大値と最小値との差の効果量を計算することで、d=0.4 が長期トレンド評価の基準値として妥当か否かを検証した。これは一種のクロスバリデーションである。その結果、基準値 d=0.4 への確たる反証は得られず、この基準値は、国の児童生徒の長期トレンド評価として、一定の妥当性をもつとの仮説が支持された。

A-3 相関係数

本調査研究では、r 族の効果量として、ピアソンの積率相関係数 r を利用した。これは、2 つの変数の相関の強さと方向（正負）を表す統計量である。相関係数 r は、-1 から 1 までの値をとる。正の相関は、一方の変数の値が大きくなると、他方の変数の値も大きくなるという関係である。負の相関は、一方の変数の値が大きくなると、他方の変数の値が小さくなるという関係である。絶対値にして 1 に近づくほど、相関は強くなる。2 変数の相関の強さを、相関係数ごとに分布図で可視化すると、図 1-7 になる。

---

[23] 大久保・岡田(2022), p.26.
[24] Cf., Cohen, J. (1988), p.82.
[25] 大久保・岡田(2022), p.94.
[26] 大久保・岡田(2022), p.95.
[27] Cf., 大久保・岡田(2022), p.95.
[28] Cf., ハッティ, J. (2018), p.93.
[29] Cf., Hattie, J. (2023), p.23.
[30] Cf., 田端(2023), pp.94-100.

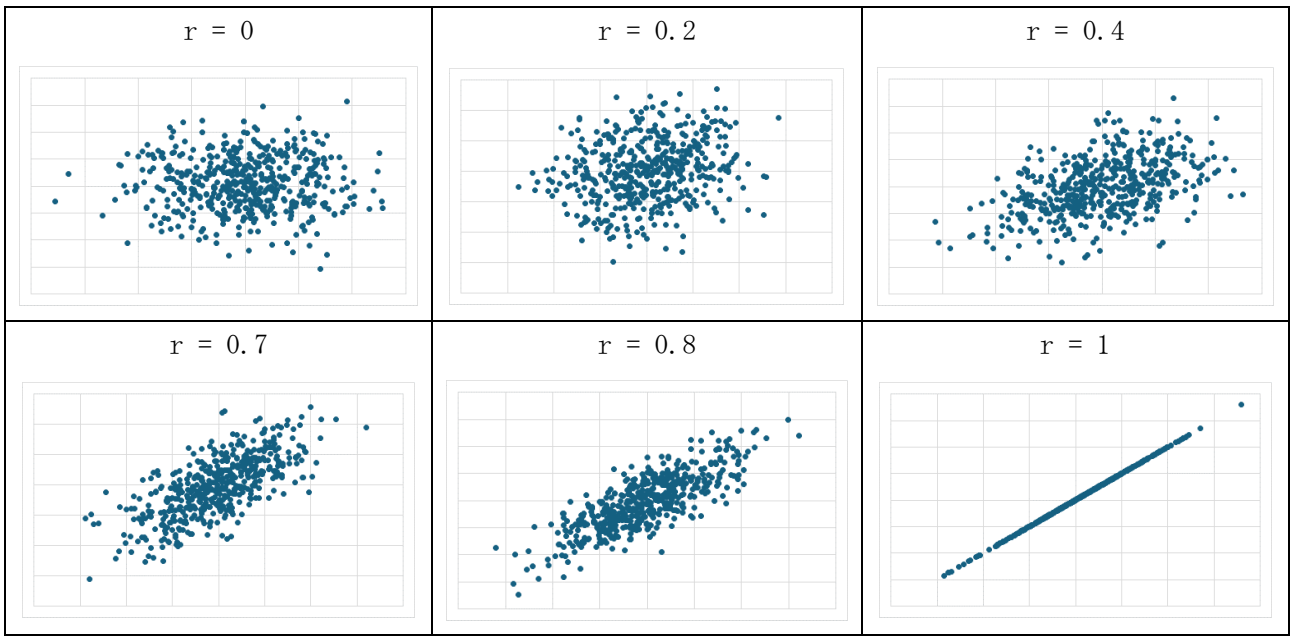


図 1-7：相関係数と散布図の対応関係

相関係数も基準値をどう定めるかで、データの評価や解釈が分かれる。しばしば絶対値にして、|r|≦0.2 は「ほとんど相関がない」、0.2＜|r|≦0.4 は「弱い相関がある」、0.4＜|r|≦0.7 は「中程度の相関がある」、0.7＜|r|≦1.0 は「強い相関がある」とされる[31]。この基準値は、図 1-7 の散布図からも直感的に妥当と思われる。

相関係数 r の一般的な基準値を、効果量 d のコーエンの基準値と対応させると、表 1-5 になる。

表 1-5：相関係数 r と効果量 d の基準値の対応表

一般的な基準値	r	d	コーエンの基準値
大	0.70	1.95	
中	0.41	0.90	
	0.37	0.80	大
	0.24	0.50	中
小	0.21	0.42	
	0.10	0.20	小

相関係数で小さいと評価される r=0.24 は、効果量にすると d=0.50 と中程度の効果と評価される。中程度に届かない効果量 d=0.42 は、相関係数で見積もると r=0.21 となり、弱いと評価できる最低限の相関になる。相関係数の基準値との対応で見ても、効果量 0.4 から 0.5 を、実質的な差といえる下限と評価することは穏当であろう。

[31] ウェブサイト「心理学用語の学習」「相関とは」を参照。
https://psychologist.x0.com/terms/617.html

A-4 帰無仮説検定

何らかの働きかけの効果の有無を統計的に判定する最も一般的な検定は、「帰無仮説検定」である。例えば、2群のテスト結果の平均に「有意差」があるかないかを判定するt検定も、帰無仮説検定の一つである。帰無仮説検定は、一般に次の手順で実施される[32]。①帰無仮説H_0と対立仮説H_1をたてる。②検定統計量（T）と分布を決める。③有意水準αを決定し、棄却域を決める。④データを取得し検定統計量（T）を算出する。⑤検定統計量が棄却域にあれば仮説H_0を棄却しH_1を採択、そうでなければH_0を採択。帰無仮説H_0は、「・・・の効果はない」とか「・・・の差はない」といった否定形をとる。有意水準αが基準値であり、慣例上有意水準5%（$p<0.05$）とされることが多い。この基準値はあくまで慣例であり、絶対的客観的な基準でないことは、繰り返し強調しておきたい。検定統計量を計算し、それが棄却域にあれば、帰無仮説を棄却し、「有意差あり」の判定になる。このように、「5%という有意水準を分水嶺として『2値的』な判断を行うことが帰無仮説検定の本質」[33]である。

有意差の有無を判定する点に帰無仮説検定の魅力とわかりやすさがあるが、この検定には幾つかの問題点がある。その一つが、先に述べたサンプル数の影響である。全国学力・学習状況調査やTIMSSやPISAといった大規模調査の場合、t検定で判定すると、ほぼすべての差が有意になってしまう。それゆえ、本調査研究では、帰無仮説検定は利用しなかった。あくまで参考値として、比較的大きなサンプル数を仮定し、「大きなサンプルでも『有意差なし』になるほど差は小さい」といった追加情報にとどめた。

---

[32] 大久保・岡田(2022), p. 23.

[33] 大久保・岡田(2022), p. 26.

２．理科の教科に関する調査の男女比較

―国内および国際学力調査（TIMSS、PISA）における検討―

２．１．理科の平均値の男女比較―国内調査での検討―

リサーチクエスチョン：

全国学力・学習状況調査の教科に関する調査で、理科の男女の平均値に実質的な差はあるか？

平成 24、27、30、令和 4 年度の小学 6 年（以下「小 6」）と中学 3 年（以下「中 3」）の男女の平均値差を検証する。

分析結果：

1. 4 か年（平成 24、27、30、令和 4 年度）計 7 回（平成 27 年度小 6 の調査については性別情報なし）の調査全てで、女子の方が男子より平均正答率が高い。
2. 4 か年の調査で、わが国の小学 6 年生（児童）および中学 3 年生（生徒）の理科の男女の平均値には「実質的な差はない」。
3. 4 か年計 7 回の調査全てで、女子の方が男子より標準偏差が小さく、得点のばらつきが小さい。

表 2-1：4 か年の理科調査における平均値の男女比較

令和 4 年度	データ件数 (N)	理科平均正答率 (mean)	標準偏差 (S.D.)	ポイント差（女子-男子）	平均値差効果量 (d)
小 6 女子	483,766	66.00	21.26	5.77	0.26
小 6 男子	502,219	60.23	23.14		
中 3 女子	447,729	50.35	18.83	1.16	0.06
中 3 男子	470,029	49.19	19.93		

平成 30 年度	データ件数	理科平均正答率	標準偏差	ポイント差（女子-男子）	平均値差効果量
小 6 女子	516,125	62.28	19.00	3.77	0.19
小 6 男子	532,520	58.51	20.81		
中 3 女子	487,895	65.90	20.25	2.29	0.11
中 3 男子	504,380	63.61	22.90		

平成 27 年度	データ件数	理科平均正答率	標準偏差	ポイント差（女子-男子）	平均値差効果量

小 6 女子	n/a	n/a	n/a	n/a	n/a
小 6 男子	n/a	n/a	n/a		
中 3 女子	517,001	54.82	21.86	2.59	0.11
中 3 男子	583,894	52.23	23.98		

※ 平成 27 年度小 6 調査には、「性別」の変数無し。

平成 24 年度	データ件数	理科平均正答率	標準偏差	ポイント差（女子-男子）	平均値差効果量
小 6 女子	128,392	62.96	18.54	3.19	0.16
小 6 男子	133,506	59.77	20.79		
中 3 女子	208,985	52.43	19.68	0.89	0.04
中 3 男子	214,538	51.54	21.61		

分析方法：

① R4 は個票データから性別と理科の平均正答率を抽出。AB 問題の H30、27、24 は AB 問題の正答数を合算し、全問題数で割って 100 をかける。

② SPSS にて「性別（０：不明/１：男子/２：女子)」によりファイルを分割。

③ グループごとのデータ件数（度数）と平均正答率と標準偏差を算出。

④ DS-EFA の「平均値差分析検定システム（AVES)」により効果量 d を算出。

⑤ 効果量（Cohen の d 値）の基準値としては、効果量 0.40 未満（d<0.40）を「実質的な差はない」と評価し、効果量 0.40 以上 0.50 未満（0.40≦d＜0.50）を「小さいが実質的な差がある」と評価する。

⑥ 平成 24 年度調査は、悉皆ではなく、抽出と希望利用方式であった。サンプル（標本）の記述統計により平均と標準偏差を計算。

２．２．国語の平均値の男女比較―国内調査での検討―

リサーチクエスチョン：

全国学力・学習状況調査の教科に関する調査で、国語の男女の平均値に実質的な差はあるか？

平成 24、27、30、令和 4 年度の小 6 と中 3 の男女の平均値差を検証する。

分析結果：

1. 4 か年（理科を実施した平成 24、27、30、令和 4 年度）計 7 回の調査全てで、女子の方が男子より平均正答率が高い。
2. 4 か年計 7 回の調査で、わが国の小学 6 年生（児童）および中学 3 年生（生徒）の国語の男女の平均値には「小さいが実質的な差がある」。7 回中 4 回の調査で、平均差効果量が 0.40 以上である。0.40 を下回る 3 回の調査でも、最低で 0.35 である。
3. 4 か年計 7 回の調査全てで、女子の方が男子より標準偏差が小さく、得点のばらつきが小さい。

表 2-2：4 か年の国語調査における平均値の男女比較

令和 4 年度	データ件数 (N)	国語平均正答率 (mean)	標準偏差 (S.D.)	ポイント差（女子-男子）	平均値差効果量 (d)
小 6 女子	484,123	69.76	21.47	8.09	0.35
小 6 男子	502,513	61.67	24.35		
中 3 女子	448,180	73.76	18.33	8.73	0.42
中 3 男子	470,294	65.03	22.30		

平成 30 年度	データ件数	国語平均正答率	標準偏差	ポイント差（女子-男子）	平均値差効果量
小 6 女子	514,475	68.58	19.46	8.15	0.40
小 6 男子	529,531	60.43	21.31		
中 3 女子	483,843	76.66	14.10	6.60	0.41
中 3 男子	498,073	70.06	17.78		

平成 27 年度	データ件数	国語平均正答率	標準偏差	ポイント差（女子-男子）	平均値差効果量
小 6 女子	n/a	n/a	n/a	n/a	n/a
小 6 男子	n/a	n/a	n/a		
中 3 女子	517,481	77.62	15.33	6.87	0.39
中 3 男子	539,083	70.75	19.28		

平成 24 年度	データ件数	国語平均正答率	標準偏差	ポイント差（女子-男子）	平均値差効果量
小 6 女子	128,752	75.56	15.74	7.55	0.43
小 6 男子	133,932	68.01	19.11		
中 3 女子	208,842	76.35	14.60	6.00	0.36
中 3 男子	214,157	70.35	18.18		

分析方法：

① 理科の場合と同様。

② 基準値としては、効果量 0.40 未満を「実質的な差はない」と評価し、効果量 0.40 以上 0.50 未

満（0.40≦d＜0.50）を「小さいが実質的な差がある」と評価する。※基準値超えの箇所を、薄い赤でハイライト

２．３．算数・数学の平均値の男女比較―国内調査での検討―

リサーチクエスチョン：

全国学力・学習状況調査の教科に関する調査で、算数・数学の男女の平均値に実質的な差はあるか？

平成 24、27、30、令和 4 年度の小 6 と中 3 の男女の平均値差を検証する。

分析結果：

1. 4 か年（理科を実施した平成 24、27、30、令和 4 年度）計 7 回の調査全てで、女子の方が男子より平均正答率が高い。
2. 4 か年計 7 回の調査で、我が国の小学 6 年生（児童）および中学 3 年生（生徒）の算数・数学の男女の平均値には「実質的な差はない」。
3. 4 か年計 7 回の調査全てで、女子の方が男子より標準偏差が小さく、得点のばらつきが小さい。

表 2-3：4 か年の算数・数学調査における平均値の男女比較

令和 4 年度	データ件数 (N)	算数・数学 平均正答率 (mean)	標準偏差 (S.D.)	ポイント差（女子-男子）	平均値差 効果量 (d)
小 6 女子	483,849	63.89	21.58	1.49	0.07
小 6 男子	502,201	62.40	23.70		
中 3 女子	447,605	52.10	24.90	0.27	0.01
中 3 男子	469,751	51.83	26.64		

平成 30 年度	データ件数	算数・数学 平均正答率	標準偏差	ポイント差（女子-男子）	平均値差 効果量
小 6 女子	514,165	59.52	21.69	1.60	0.07
小 6 男子	529,057	57.92	23.71		
中 3 女子	482,098	62.37	21.08	1.65	0.07
中 3 男子	495,487	60.72	23.35		

平成 27 年度	データ件数	算数・数学 平均正答率	標準偏差	ポイント差（女子-男子）	平均値差 効果量
小 6 女子	n/a	n/a	n/a	n/a	n/a

小 6 男子	n/a	n/a	n/a		
中 3 女子	517,061	58.83	21.00	0.81	0.04
中 3 男子	538,796	58.02	23.27		

平成 24 年度	データ件数	算数・数学 平均正答率	標準偏差	ポイント差（女子-男子）	平均値差効果量
小 6 女子	128,605	65.86	18.00	0.26	0.01
小 6 男子	133,745	65.60	19.72		
中 3 女子	208,876	59.73	21.18	0.52	0.02
中 3 男子	214,315	59.21	23.04		

分析方法：

① 理科の場合と同様。

② 基準値としては、効果量 0.40 未満を「実質的な差はない」と評価し、効果量 0.40 以上 0.50 未満（0.40≦d＜0.50）を「小さいが実質的な差がある」と評価。

２．４．理数教科の平均値の男女比較―国際調査 TIMSS2019 での検討―

リサーチクエスチョン：

国内学力調査の男女比較の結果は、国際学力調査の男女比較の結果と整合的か？TIMSS2019 の中 2（G8）理数教科の日本の男女差は、国内調査と整合的か、また国際比較するとどうなるか？

分析結果：

1. 理科と数学とも、日本の男女スコアは、男子が女子よりわずかに高い。その差は棒グラフ[34]で見る限り、ほとんど認められない。「男女の理数学力平均に実質的な差はない」という、国内調査結果を利用した本調査研究結果と TIMSS2019 中 2 の男女平均差とは整合的である。
2. 抽出した 6 か国では、米国とフィンランドは、理科も数学も、女子が男子よりスコアが高い。
3. 英国は、理科で、女子が男子よりスコアが高い。

分析方法：

TIMSS の実施者である IEA（International Association for the Evaluation of Educational Achievement:国際教育到達度評価学会）が公開している分析ソフト IDB アナライザー[35]、およびそれが橋渡しする IBM 社 SPSS を利用した。

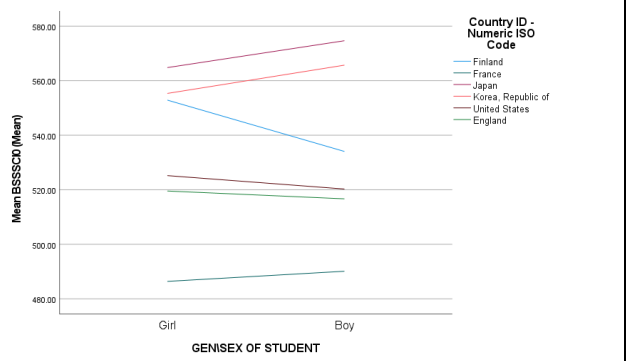
PV's（Plausible Values：推算値）を利用し、生徒重みづけ（Student Weight）を行った。

[34] 棒グラフの上の線グラフは、縦軸が 480 以上にクローズアップされているため、差が大きく見えるが、縦軸の起点を 0 にする棒グラフでは、男女のスコアは僅差である。

[35] https://www.iea.nl/data-tools/tools

国際比較のため、任意に、フィンランド、フランス、日本、韓国、米国、英国を抽出した。
今回は、中 2（G8：第 8 学年）のみの分析である。

結果の可視化：



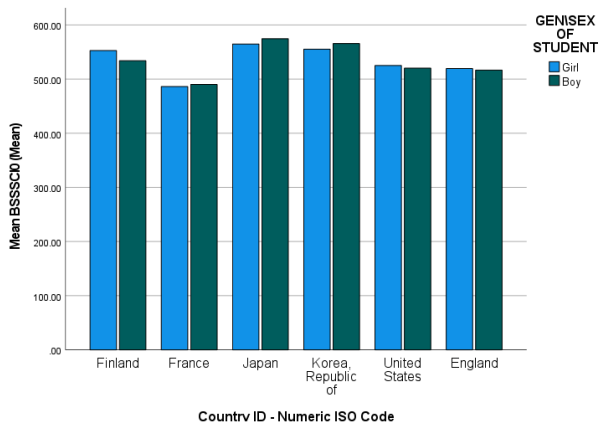
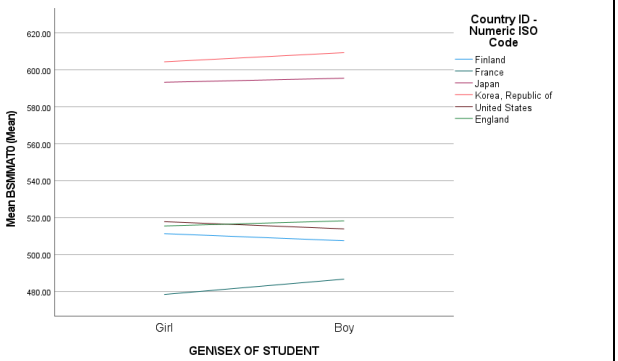


図 2-1：TIMSS2019_中 2_理科



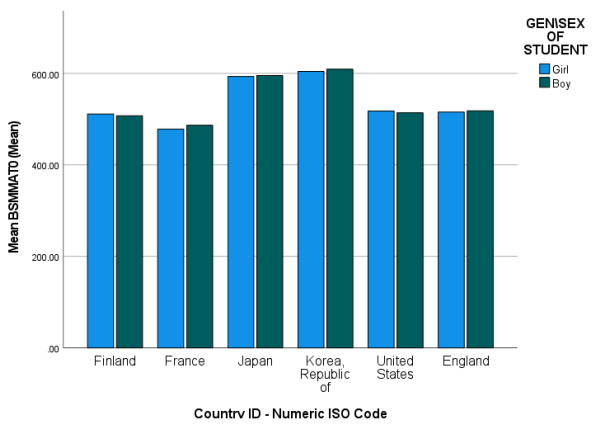


図 2-2：TIMSS2019_中 2_数学

２．５．科学的リテラシー・数学的リテラシー・読解力の平均値の男女比較 —国際調査 PISA2018 での検討—

リサーチクエスチョン：

国内学力調査の男女比較の結果は、国際学力調査の男女比較の結果と整合的か？PISA2018 の 15 歳（日本では高校 1 年生）の科学的リテラシー・数学的リテラシー・読解力[36]の日本の男女差は、国内調査と整合的か、また国際比較するとどうなるか？

分析結果：

1. 科学的リテラシーと数学的リテラシーについては、棒グラフで見る限り、6 か国とも男女差はわずかである。国内学力調査を利用した本調査研究の男女差結果と整合的。
2. 科学的リテラシーでは、フィンランド、フランス、米国で、女子が男子よりスコアが高い。

[36] 「読解力・数学的リテラシー・科学的リテラシー」という順序が一般的であるが、本調査研究は理科に焦点化しているため、理数教科を先にした。

3. 数学的リテラシーでは、フィンランドのみ、女子が男子よりスコアが高い。
4. 読解力については、6 か国すべてで、女子が男子より高い。国内学力調査を利用した本調査研究結果と合わせると、国語力については、性差が認められる可能性が高い。今後の検証が必要である。

分析方法：

本調査研究２．４と同じ。

結果の可視化：

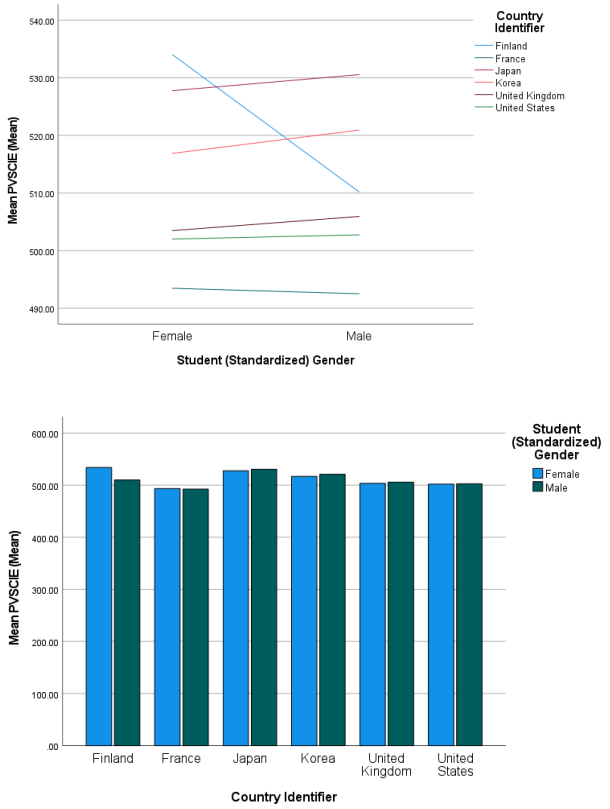


図 2-3：PISA2018_15 歳_科学的リテラシー

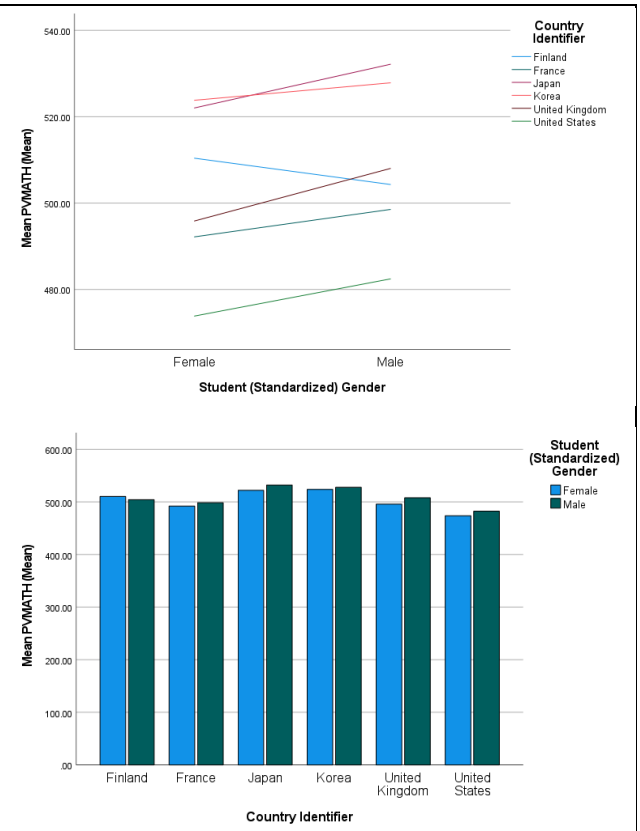
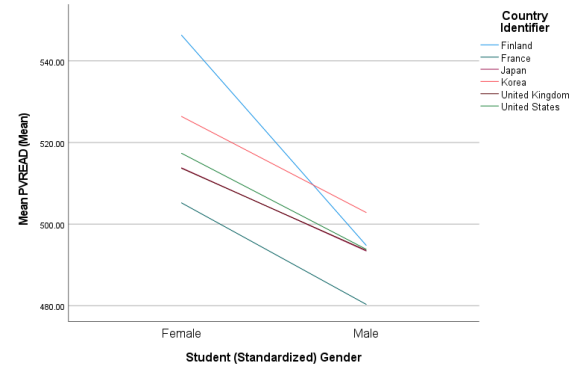


図 2-4：PISA2018_15 歳_数学的リテラシー



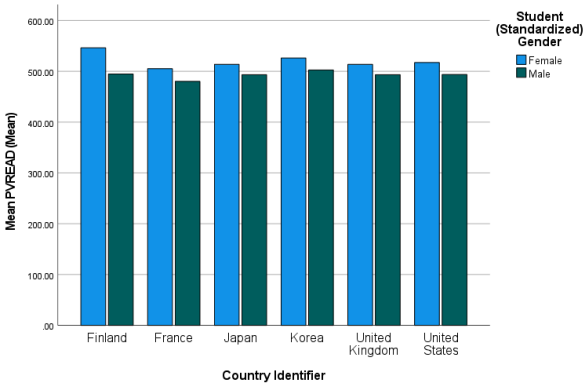


図 2-5：PISA2018_15 歳_読解力

３．教科間の相関の男女比較

リサーチクエスチョン：

理科、国語、算数・数学の教科間の相関の強さに関して、男女差はあるか？

分析結果：

1. 平成 24 年度から令和 4 年度にかけての 4 か年 7 回の調査で、同年度・同学年・同教科間の男女を比較すると、全てで男子の方が女子より相関がわずかに強い。男子の方が女子よりも、ある教科ができると別の教科もできるという傾向（またその逆：ある教科ができないと別の教科もできないという傾向）がわずかに強い。
2. 男女合わせて合計 42 か所の相関係数のうち、r＞0.7 は 34 か所あり、全体の 81％で異教科間の相関係数が 0.7 を上回っている。異教科間の相関は「大きい（強い）」。

表 3-1：4 か年における 3 教科の相関係数一覧

R4_小 6 女子	国	数	理
国	1		
数	0.68	1	
理	0.69	0.72	1
R4_小 6 女子平均：			0.70

R4_小 6 男子	国	数	理
国	1		
数	0.73	1	
理	0.73	0.76	1
R4_小 6 男子平均：			0.74

R4_中 3 女子	国	数	理
国	1		
数	0.64	1	
理	0.63	0.71	1
R4_中 3 女子平均：			0.66

R4_中 3 男子	国	数	理
国	1		
数	0.71	1	
理	0.68	0.73	1
R4_中 3 男子平均：			0.71

H30_小 6 女子	国	数	理
国	1		
数	0.73	1	
理	0.66	0.69	1
H30_小 6 女子平均：			0.69

H30_小 6 男子	国	数	理
国	1		
数	0.76	1	
理	0.69	0.73	1
H30_小 6 男子平均：			0.73

H30_中 3 女子	国	数	理
国	1		
数	0.77	1	
理	0.76	0.82	1
H30_中 3 女子平均：			0.78

H30_中 3 男子	国	数	理
国	1		
数	0.80	1	
理	0.80	0.83	1
H30_中 3 男子平均：			0.81

H27_小 6 女子	国	数	理
国	1		
数	n/a	1	
理	n/a	n/a	1

H27_小 6 男子	国	数	理
国	1		
数	n/a	1	
理	n/a	n/a	1

H27_中 3 女子	国	数	理
国	1		
数	0.77	1	
理	0.74	0.83	1
H27_中 3 女子平均:			0.78

H27_中 3 男子	国	数	理
国	1		
数	0.80	1	
理	0.77	0.84	1
H27_中 3 男子平均:			0.80

H24_小 6 女子	国	数	理
国	1		
数	0.75	1	
理	0.71	0.74	1
H24_小 6 女子平均:			0.73

H24_小 6 男子	国	数	理
国	1		
数	0.79	1	
理	0.76	0.77	1
H24_小 6 男子平均:			0.77

H24_中 3 女子	国	数	理
国	1		
数	0.75	1	
理	0.72	0.81	1
H24_中 3 女子平均:			0.76

H24_中 3 男子	国	数	理
国	1		
数	0.79	1	
理	0.76	0.82	1
H24_中 3 女子平均:			0.79

分析方法：

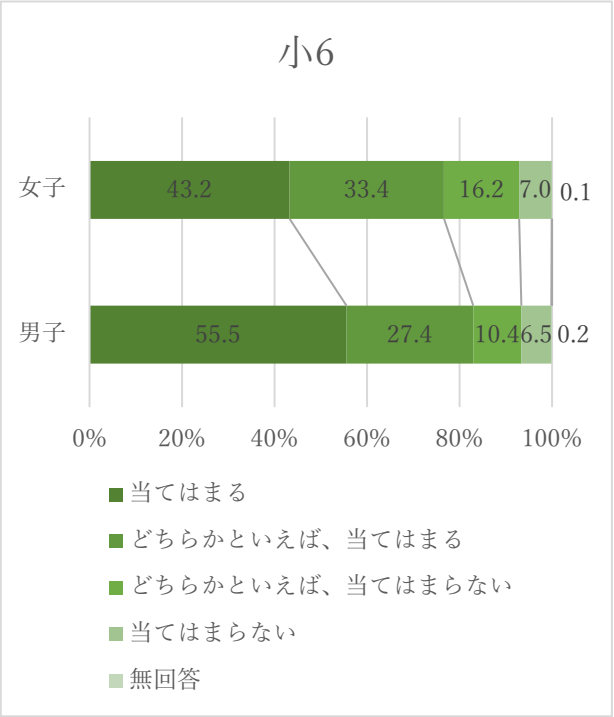
① SPSS にて男女別の相関行列を計算。

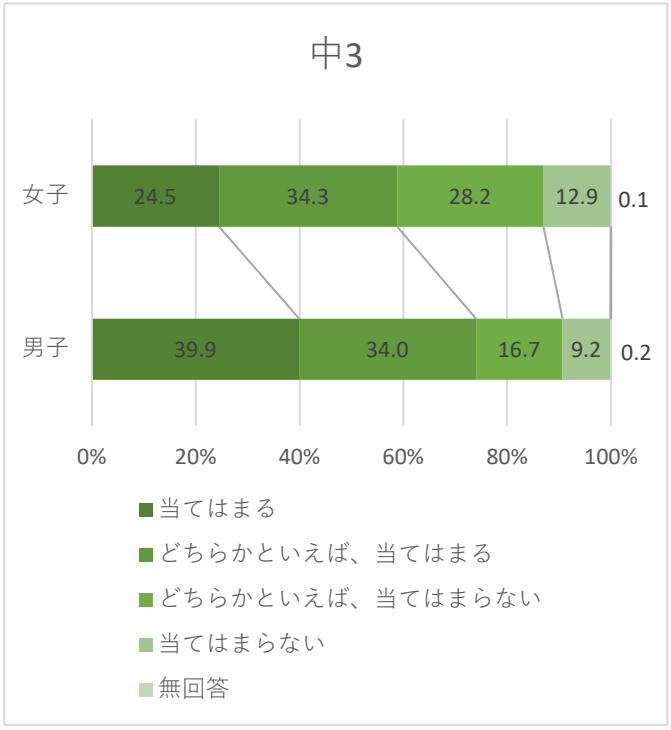
４．教科に関する質問紙調査の回答傾向の男女比較

４．１．令和４年度、理科の学習に対する興味・関心や授業の理解度等の男女比較

※　国語、数学、理科の質問番号等の表の右端の列にある数値は、質問項目と教科との相関係数ｒを示し、上段が小６、下段が中３の相関係数である。国立教育政策研究所が公開している相関係数の転記である[37]。

質問番号	質問事項	国語	数学	理科
小中（６１）	理科の勉強は好きですか	0.043	0.052	0.123
		0.102	0.179	0.253





女子は男子より理科の勉強が好きと回答する割合が小さく、その割合は小６より中３の方がいっそう小さい。

男女ともに、理科の勉強が好きとの回答割合は、中３は小６より小さい。

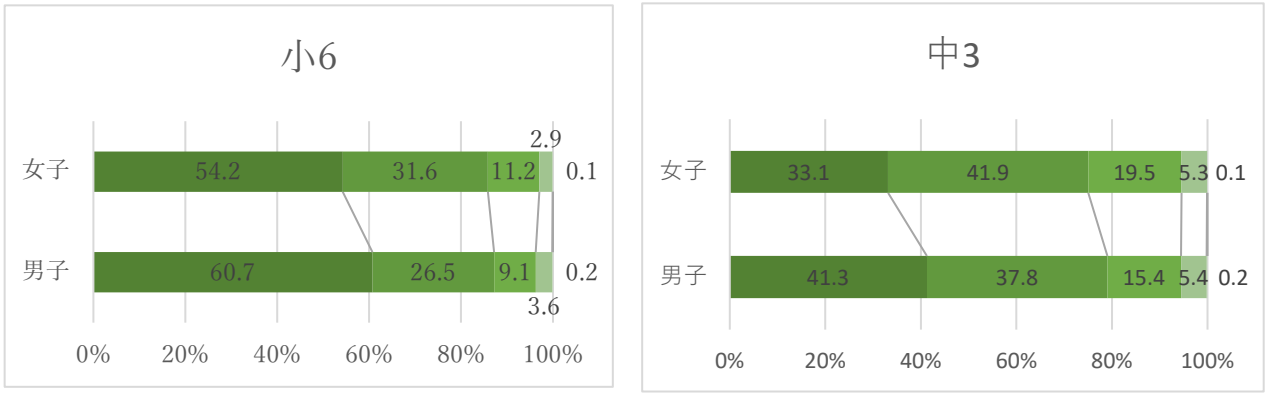
この項目の小６女子は平均 3.13 で標準偏差 0.93、男子は平均 3.32 で標準偏差 0.90 であり、差分効果量 d=0.21 となる。中３女子は平均 2.70 で標準偏差 0.98、男子は平均 3.05 で標準偏差 0.97 であり、差分効果量 d=0.36 となる。（※）

---

[37] 「全国学力・学習状況調査」のウェブページから「調査結果資料」に進み、「（３）相関係数、クロス集計表」から「相関係数（児童生徒質問紙-教科）全国【表】」をクリックすると、相関係数を記載したエクセルファイルがダウンロードされる。
【小学校版】https://www.nier.go.jp/22chousakekkahoukoku/factsheet/primary.html
【中学校版】https://www.nier.go.jp/22chousakekkahoukoku/factsheet/middle.html
［2024/3/13 最終閲覧］

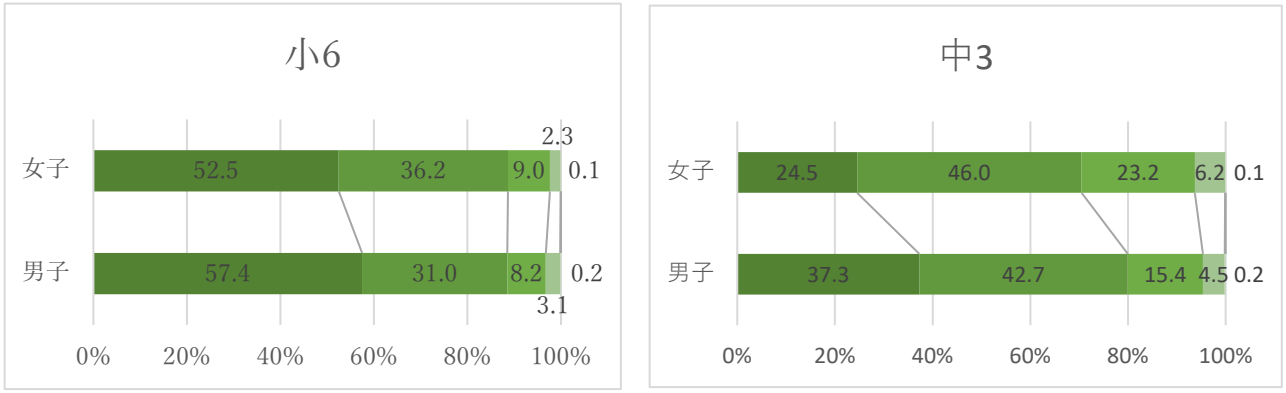
※　回答選択肢「当てはまる」に 4、「どちらかといえば、当てはまる」に 3、「どちらかといえば、当てはまらない」に 2、「当てはまらない」に 1、無回答や 99 を削除して計算した。

小中（６２）	理科の勉強は大切だと思いますか	0.094	0.100	0.137
		0.131	0.184	0.206



小 6 でも中 3 でも、女子は男子より理科の勉強を大切と思う（「当てはまる」）回答割合が小さい。
男女ともに、理科の勉強が大切と思う割合は、小 6 よりも中 3 の方が小さい。

小中（６３）	理科の授業の内容はよく分かりますか	0.178	0.177	0.236
		0.174	0.249	0.284



女子は男子よりも、理科の内容がよく分かる（「当てはまる」）との回答割合が小さい。
理科の内容がよくわかると回答する割合は、男女とも、小 6 より中 3 の方がいっそう小さい。

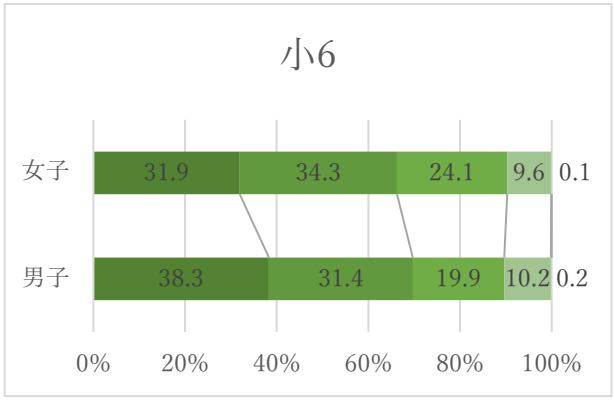
＃【理科に関する質問紙項目の分析結果の解釈 01】

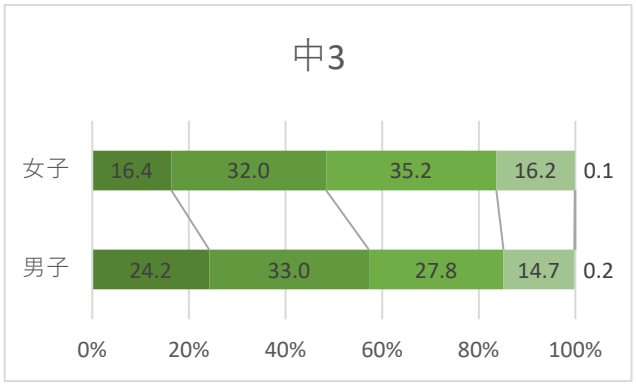
1. 理科が好き、大切と思う、授業が分かるの点で、小 6 中 3 ともに、女子は男子より「当てはまる」と回答する割合が小さい。
2. これら 3 つの質問項目は、中 3 では理科学力値と弱い相関がある。男女合算で、理科が好きで、大切と思い、授業が分かるほど理科学力値が高い傾向があり、その逆でもある。

＃

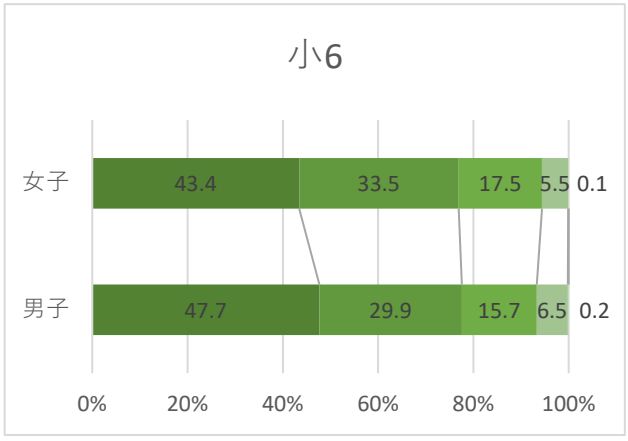
小中（６４）	理科の授業で学習したことを，普段の生活の中で活用で	0.082	0.080	0.107

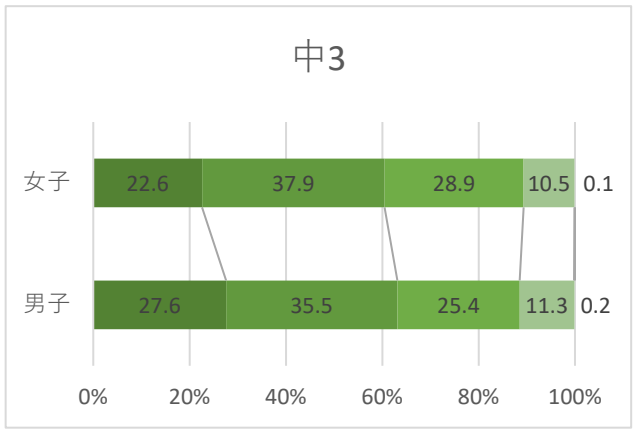
	きないか考えますか	0.106	0.170	0.198



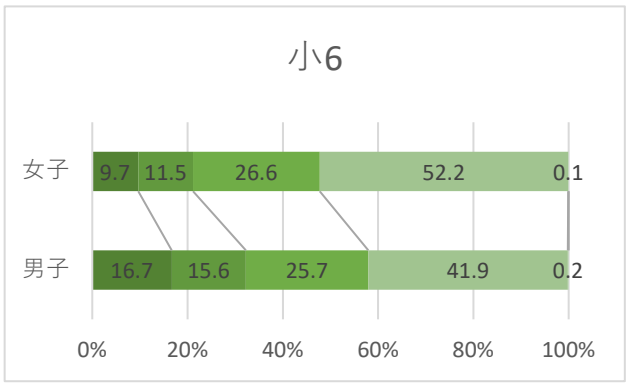


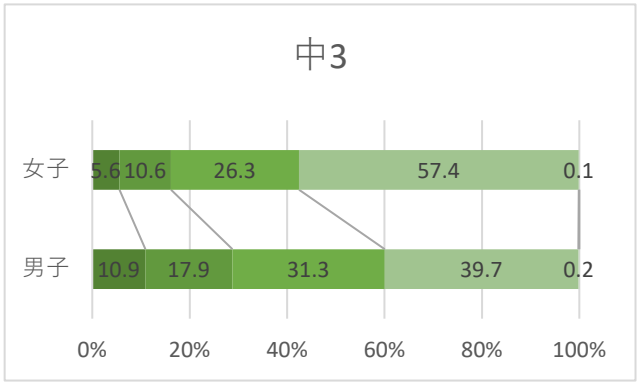
小中（６５）	理科の授業で学習したことは，将来，社会に出たときに役に立つと思いますか	0.090	0.090	0.120
		0.114	0.159	0.182





小中（６６）	将来，理科や科学技術に関係する職業に就きたいと思いますか	−0.001	0.032	0.030
		0.051	0.170	0.192





この項目の小 6 女子は平均 1.79 で標準偏差 0.99、小 6 男子は平均 2.07 で標準偏差 1.11 であり、差分効果量 d=0.27 となる。中 3 女子は平均 1.65 で標準偏差 0.88、中 3 男子は平均 2.00 で標準偏差 1.01 であり、差分効果量 d=0.37 となる。

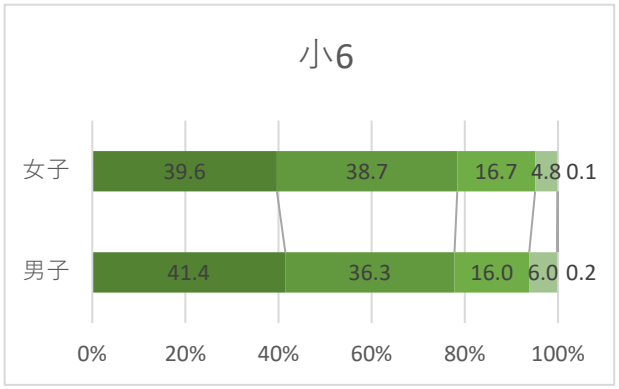
＃【理科に関する質問紙項目の分析結果の解釈 02】

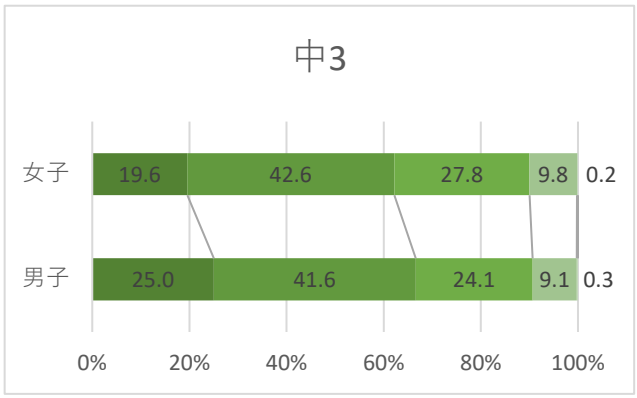
1. 理科を生活で活用したり、将来社会で役立つと感じたり、科学技術職に就きたいに関し、小６中３ともに、女子は男子より「当てはまる」と回答する割合が小さい。

・ これらの動機が小さいにもかかわらず、女子は男子より理科平均正答率が高い。

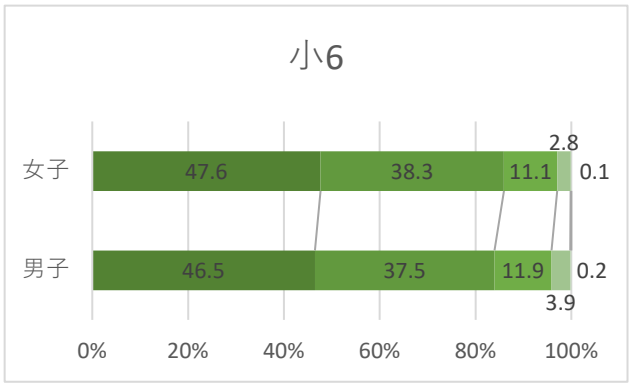
#

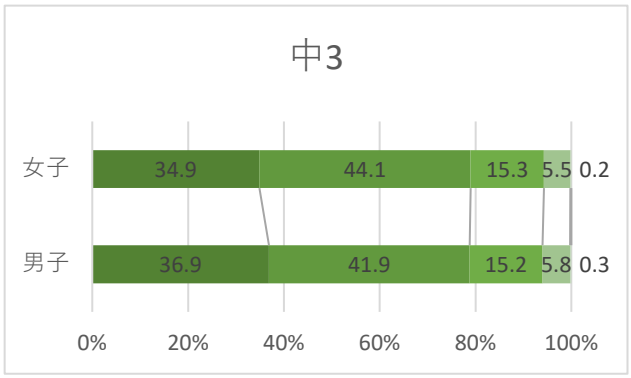
小中（６７）	理科の授業では，自分の予想をもとに観察や実験の計画を立てていますか	0.182	0.176	0.202
		0.160	0.195	0.199



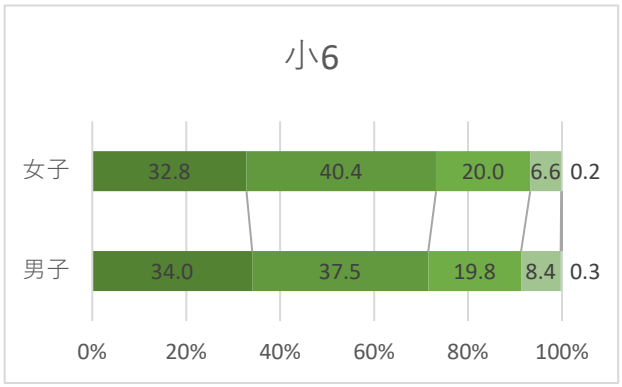


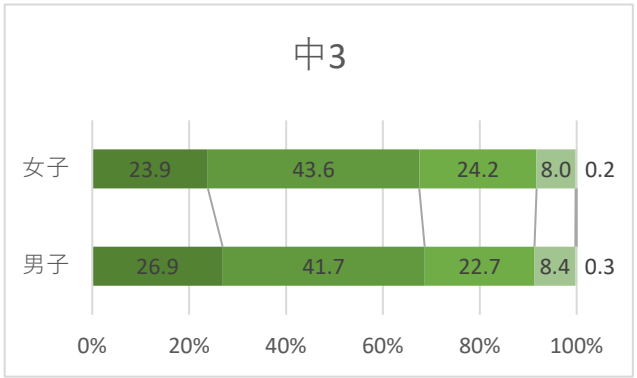
小中（６８）	理科の授業で，観察や実験の結果から，どのようなことが分かったのか考えていますか	0.239	0.231	0.265
		0.255	0.283	0.292





小中（６９）	理科の授業で，観察や実験の進め方や考え方が間違っていないかを振り返って考えていますか	0.147	0.138	0.158
		0.199	0.232	0.236





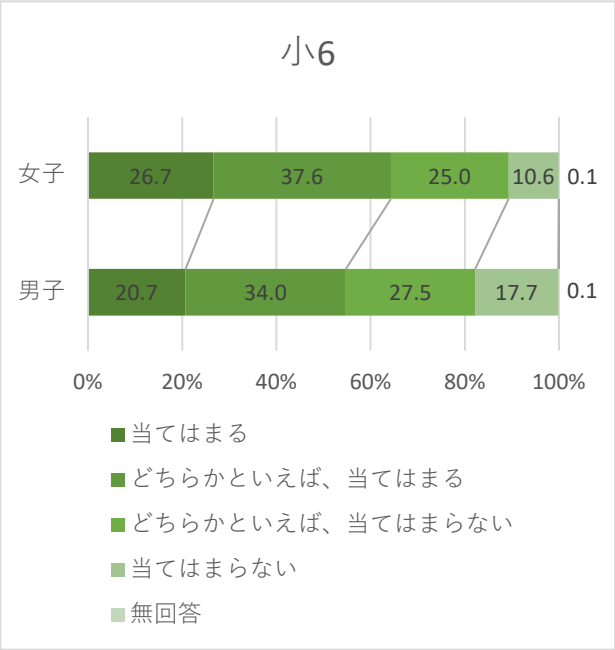
#【理科に関する質問項目の分析結果の解釈 03】

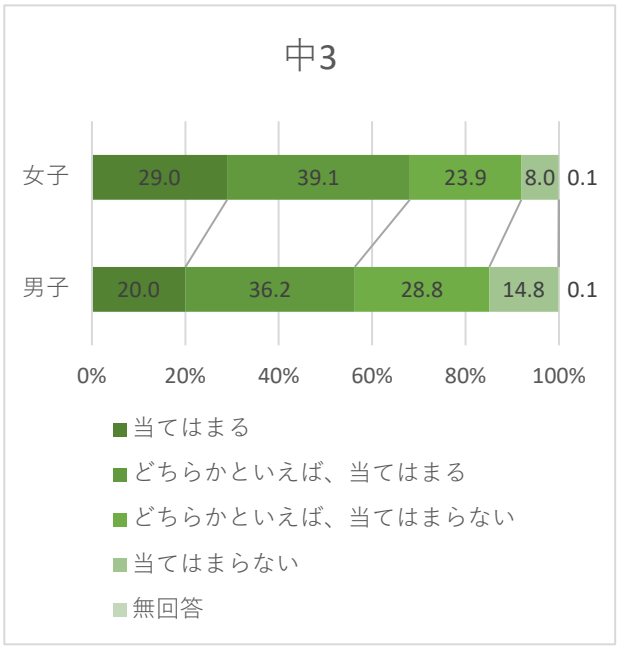
1. 予測にもとづき観察や実験の計画を立てる、観察や実験結果から推測する、観察や実験に間違いがないかチェックする論理的・科学的スキルに関して、小 6 中 3 ともに男女に顕著な差はない。

#

４．２．令和４年度、国語の学習に対する興味・関心や授業の理解度等の男女比

質問番号	質問事項	国語	数学	理科
小中（４９）	国語の勉強は好きですか	0.214	0.101	0.158
		0.150	0.013	0.072

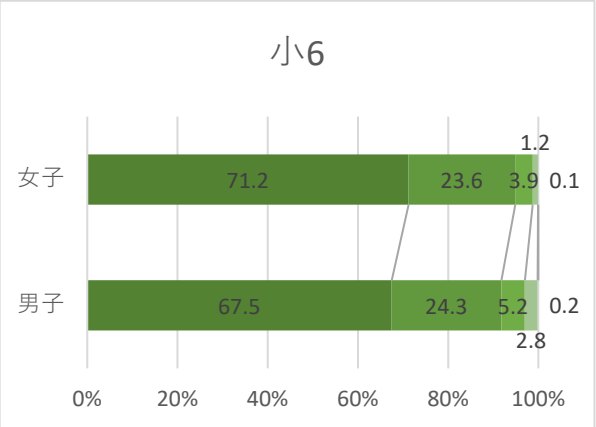


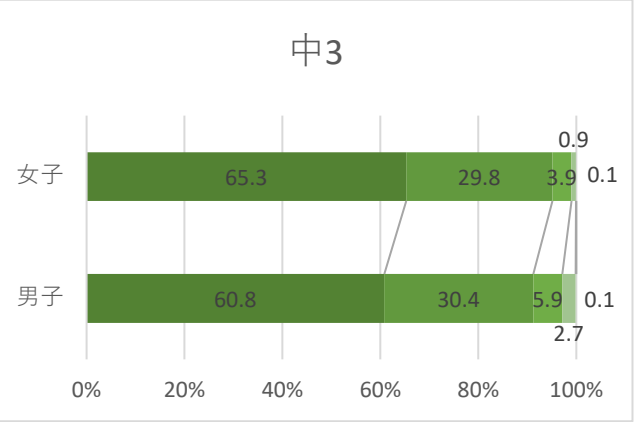


女子は男子よりも、国語の勉強が好き（「当てはまる」）と回答する割合が大きい。

理科とは異なり、国語では、男女とも、中３は小６よりも好き回答する割合が（「当てはまる」の割合も、「どちらかといえば当てはまる」の割合も）大きい。

小中（５０）	国語の勉強は大切だと思いますか	0.206	0.172	0.193
		0.121	0.069	0.072

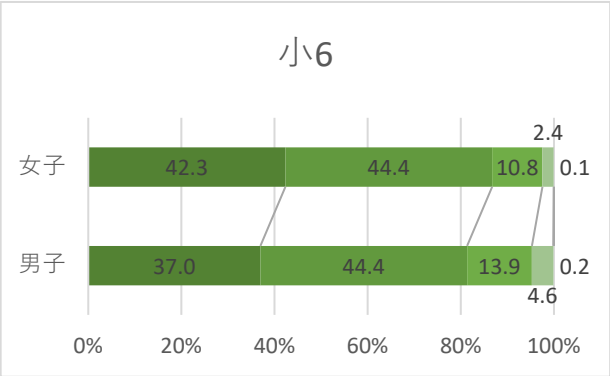


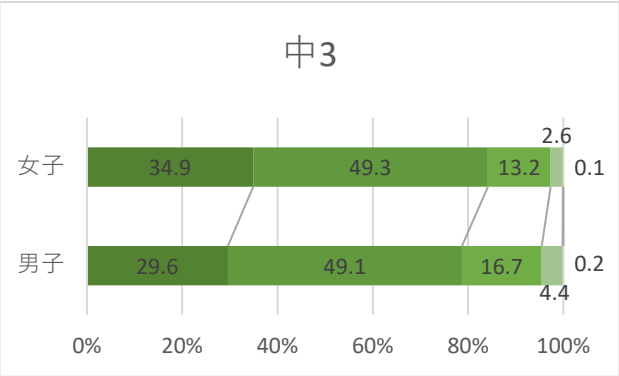


小中ともに、女子は男子より、国語の勉強を大切だと思うと回答する割合が大きい。

小中（５１）	国語の授業の内容はよく分かりますか	0.336	0.272	0.296

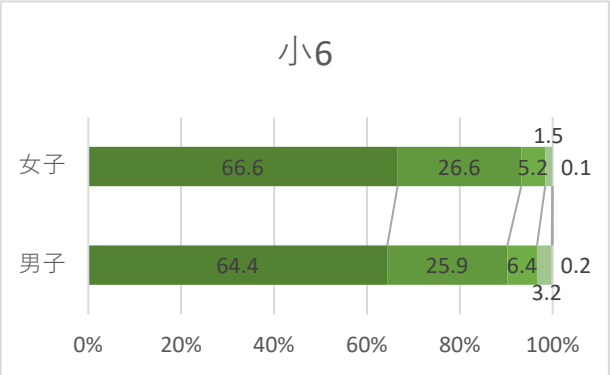
		0.259	0.187	0.204

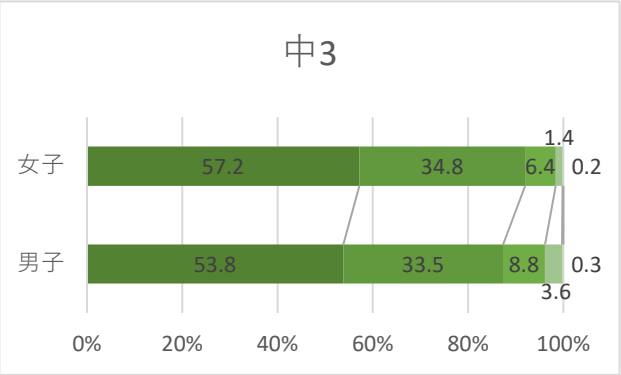




小中ともに、女子は男子より、国語の授業の内容が分かると回答する割合が大きい。

小中（５２）	国語の授業で学習したことは，将来，社会に出たときに役に立つと思いますか	0.160	0.138	0.156
		0.086	0.037	0.048





小中ともに、女子は男子より、国語の授業で学習したとき将来社会で役立つと回答する割合が大きい。

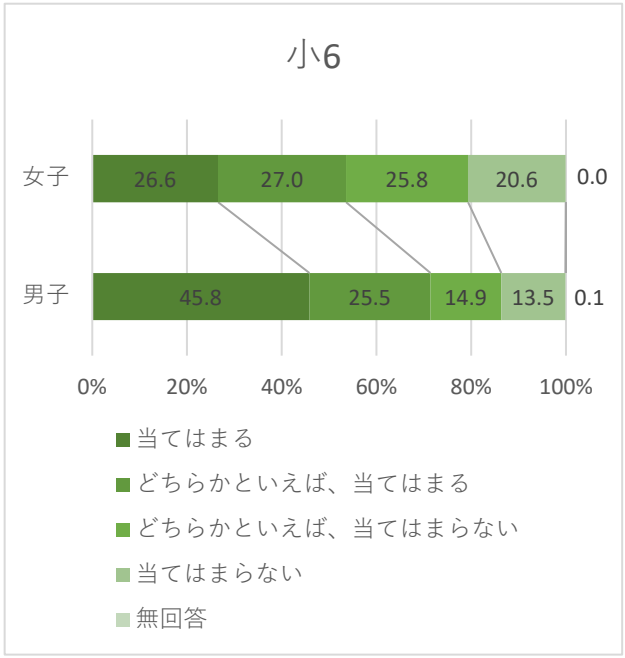
＃【国語に関する質問項目の分析結果の解釈】

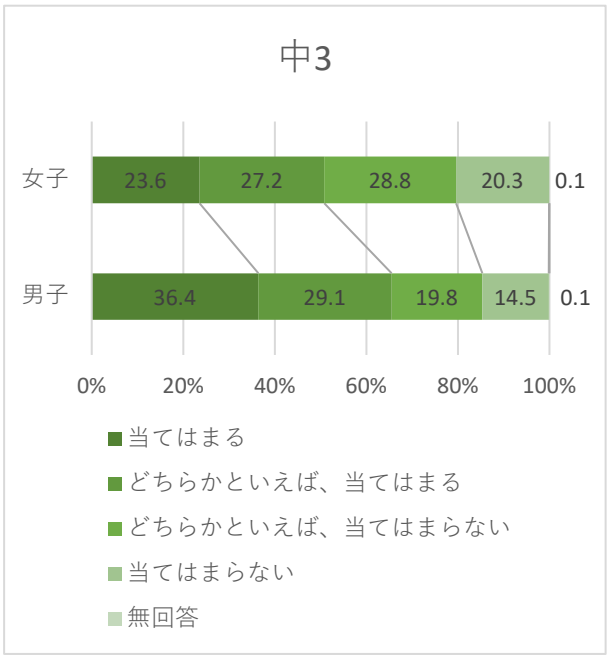
1. 小６中３とも、理科とは逆に、全ての項目で女子の方が男子より肯定的な回答割合が大きい。
2. 国語が好きで、大切だと思い、授業が分かり、社会で役立つと思う傾向が強いことが、小６中３とも、女子の平均正答率を男子より高く押し上げている可能性がある。

＃

４．３．令和４年度、算数・数学の学習に対する興味・関心や授業の理解度等の男女比

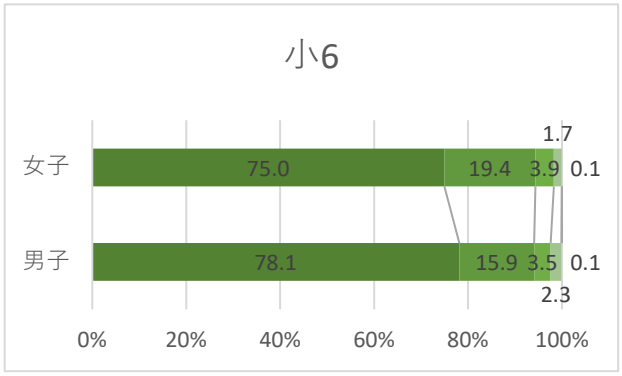
質問番号	質問事項	国語	数学	理科
小中（５３）	算数の勉強は好きですか	0.188	0.340	0.237
		0.157	0.377	0.244

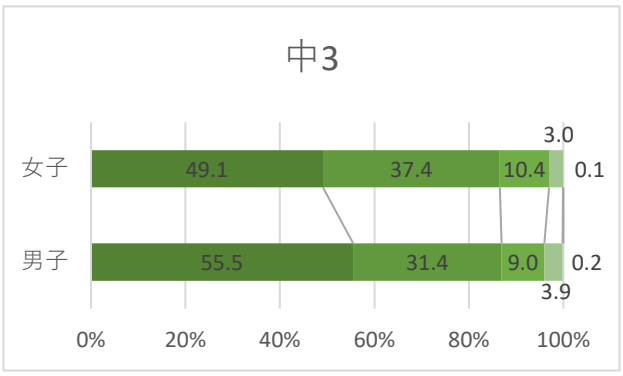




女子は男子よりも算数・数学の勉強が好きと回答する割合が小さい。

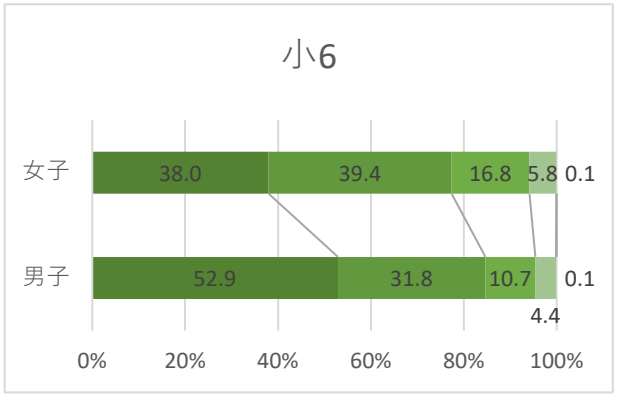
小中（５４）	算数の勉強は大切だと思いますか	0.185	0.220	0.201
		0.122	0.186	0.148

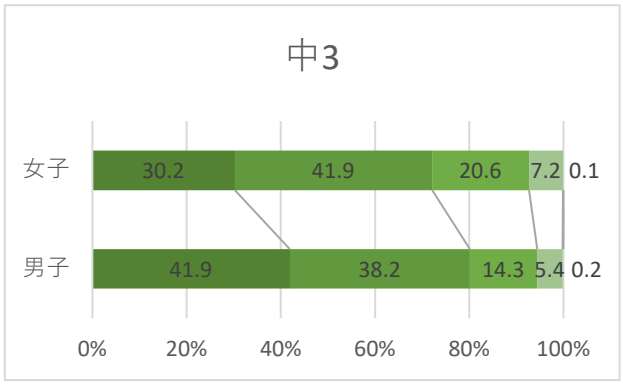




女子は男子よりも算数・数学の勉強を大切と思う割合が小さい。
男女ともに、小６より中３の方が、算数・数学の勉強を大切と思う割合は小さい。

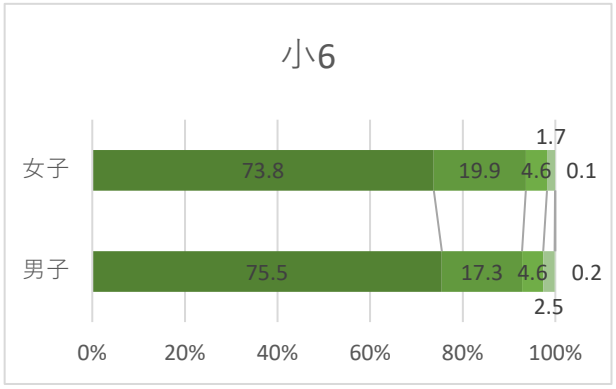
小中（５５）	算数の授業の内容はよく分かりますか	0.301	0.428	0.334
		0.230	0.415	0.288

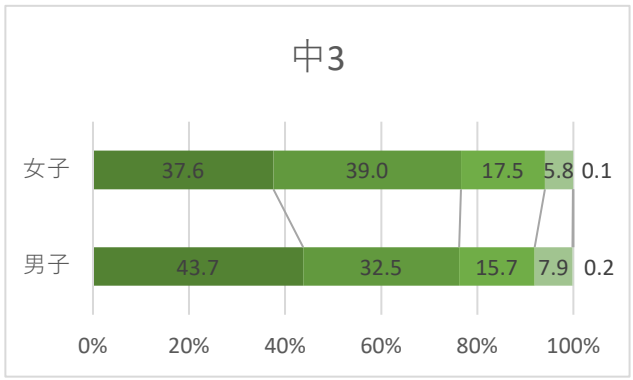




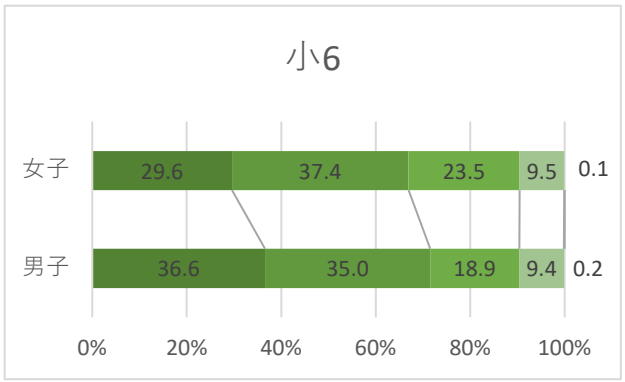
女子は男子より、算数・数学の授業の内容が分かると回答する割合が小さい。

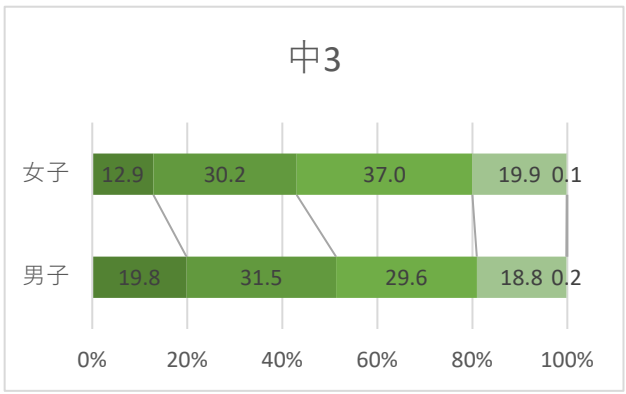
小中（５６）	算数の授業で学習したことは，将来，社会に出たときに役に立つと思いますか	0.141	0.157	0.154
		0.052	0.101	0.077



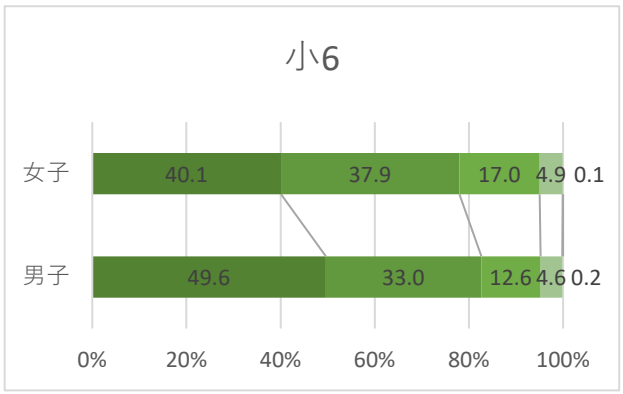


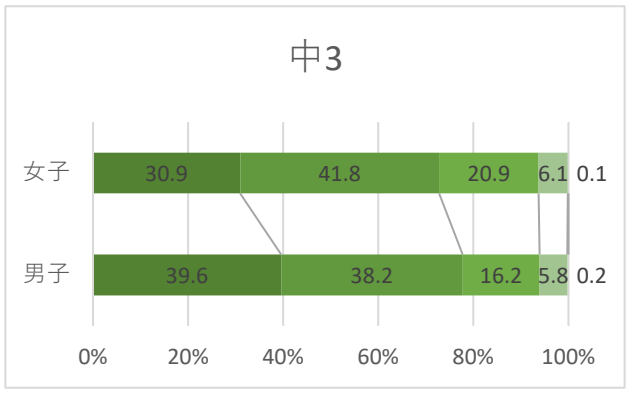
小中（５７）	算数の授業で学習したことを，普段の生活の中で活用できないか考えますか	0.137	0.179	0.144
		0.086	0.193	0.138



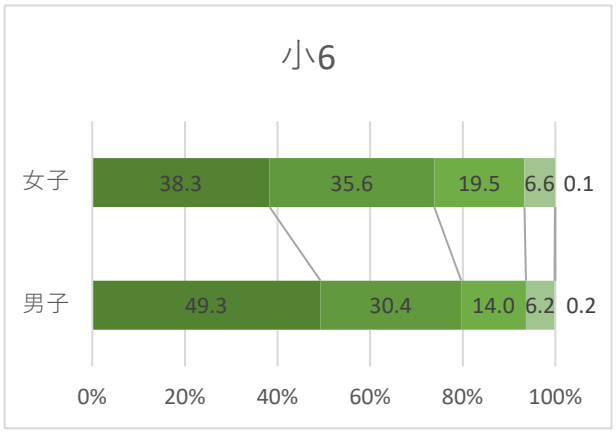


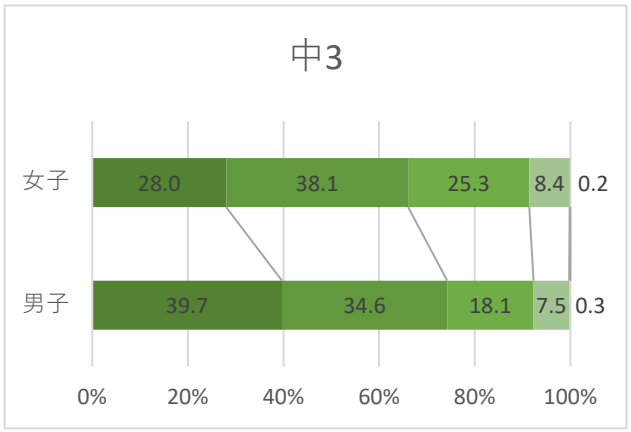
小中（５８）	算数の問題の解き方が分からないときは，あきらめずにいろいろな方法を考えますか	0.237	0.304	0.252
		0.240	0.385	0.284



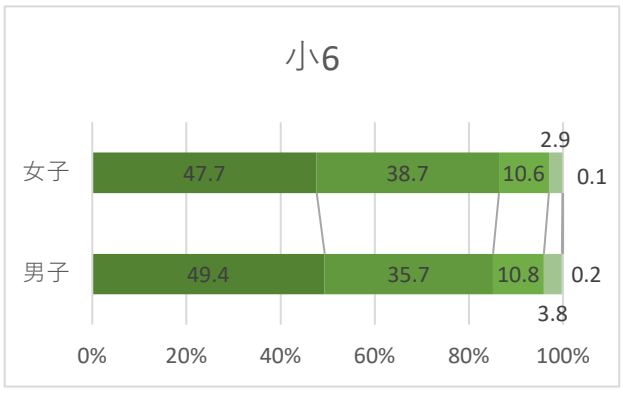


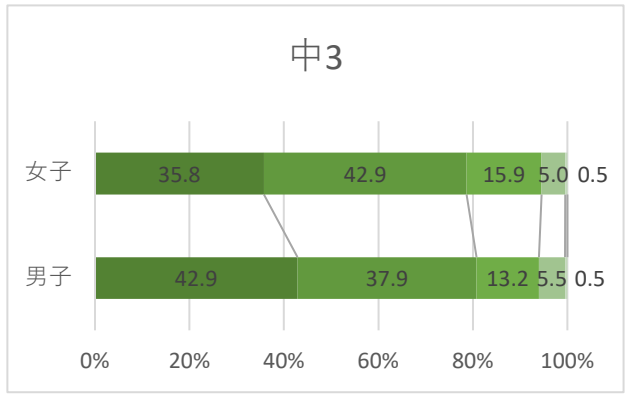
小中（５９）	算数の授業で問題を解くとき，もっと簡単に解く方法がないか考えますか	0.195	0.251	0.197
		0.166	0.294	0.211





小中（６０）	算数の授業で公式やきまりを習うとき，そのわけを理解するようにしていますか	0.274	0.306	0.273
		0.191	0.283	0.227





＃【算数・数学に関する質問項目の分析結果の解釈】

1. 理科に関する質問項目と同様に、算数・数学の勉強を大切だと思う、授業内容がよく分かる、社会に出たときに役立つと思うの項目で、女子は男子より肯定的に回答する割合が小さい傾向がうかがわれる。

2. 算数・数学の授業で学習したことを、普段の生活の中で活用できないか考える項目でも、女子は男子より肯定的に回答する割合が小さい。
3. 算数・数学の問題の解き方が分からないとき、あきらめずにいろいろな方法を考える項目でも、女子は男子より肯定的に回答する割合が小さい。
4. 算数・数学の公式や規則を習うとき、その根拠を理解する項目でも、女子は男子より肯定的に回答する割合が小さい。

#

４．４．TIMSS2019、理数教科の学習に関する質問紙調査の男女比較

リサーチクエスチョン：

４．１．や４．３．の傾向は、国際学力調査 TIMSS2019 でも見られるか？また、他の先進国と比較して、日本の男女差の傾向に特徴はあるか？

結論：

日本の小４中２ともに、４．１．や４．３．と同様の結果が見られた。

抽出した６か国で、日本と同じ傾向がうかがわれるが、その程度には差がある。

米国やフィンランドでは、女子が男子より、理科が好きな傾向や、理科に自信がある傾向がある。

指導の明確さも、おしなべて男子が女子より、理数の指導が明確と感じる程度が高いが、逆転するケース（米国やフィンランドの小４など）もある。

理数教科の価値についても、フィンランドを除く５か国で、女子は男子より、価値を見出しにくい傾向がある。日本と韓国は、他国より、男女とも、理数系に価値を見出す傾向が低い。

【理科の分析結果】

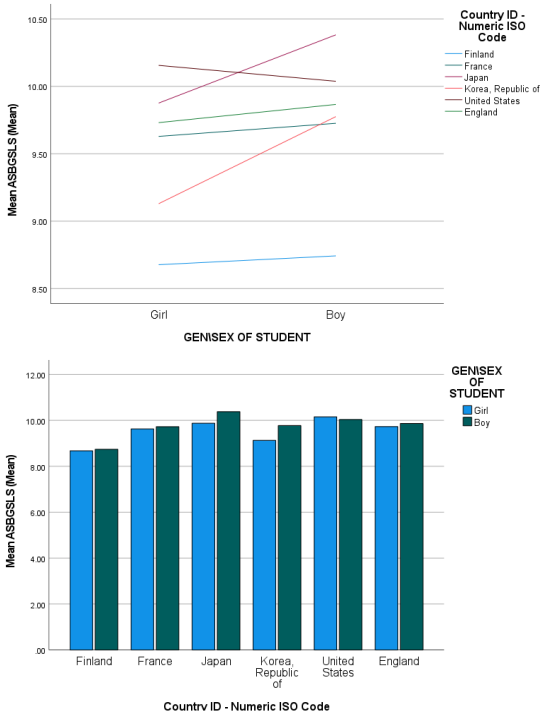


図 2-1：TIMSS_小４理科_「好き」尺度（ASBGSLS）

・日本と韓国の男女差は、他の４か国より大きい。

・米国は女子が男子より理科が好き。

・日本は男女ともに棒が長く、他の５か国より

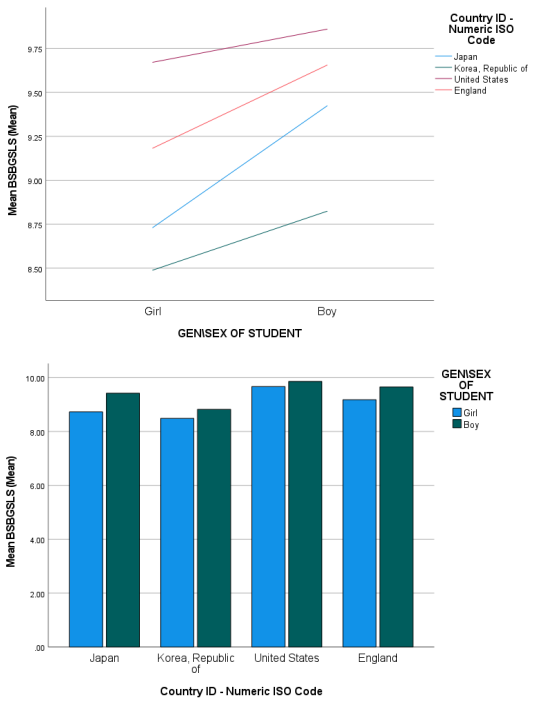


図 2-2：TIMSS_中２理科_「好き」尺度（BSBGSLS）[38]

・４か国とも、男子の方が女子より理科が好き。

[38] この項目は、フィンランドとフランスにはなかったため、４か国比較になっている。この２国は、理科の分野別に授業の明確さを質問している。

理科の学習を好む傾向にある。	

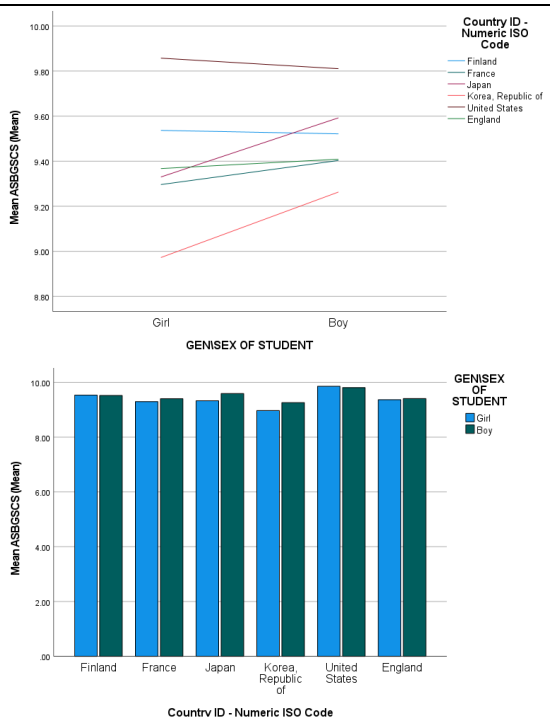


図 2-3：TIMSS_小 4 理科_「自信」尺度（ASBGSCS）

・日本と韓国の男女差は、他の 4 か国より大きい。

・フィンランドと米国は、女子の方が男子より理科に自信がある。

・米国の児童が男女ともに最も理科に自信がある。

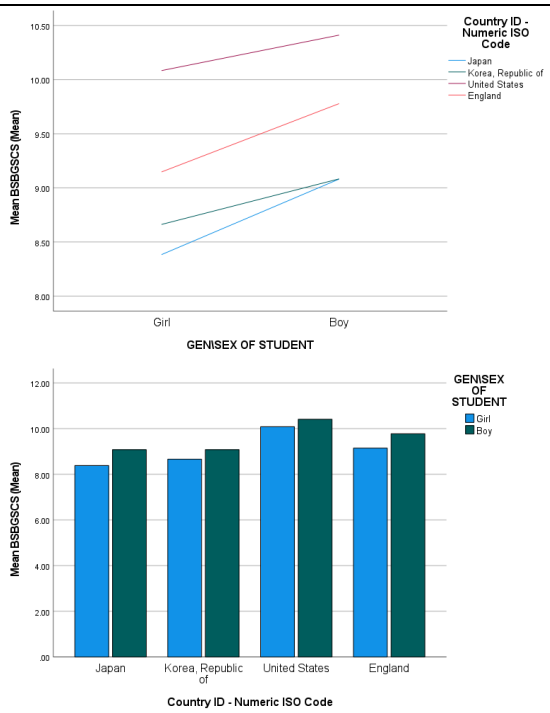
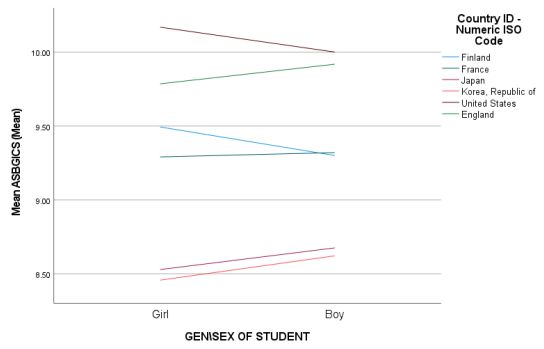
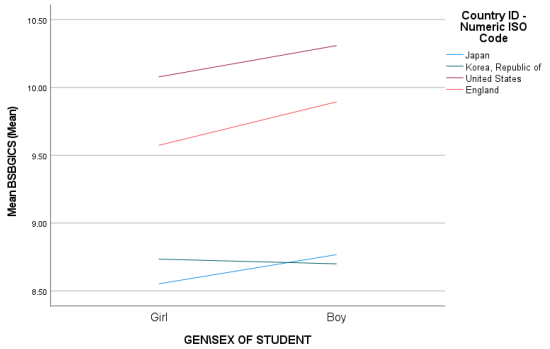


図 2-4：TIMSS_中 2 理科_「自信」尺度（BSBGSCS）[39]

・4 か国とも、男子の方が女子より理科に自信がある。

・米国が、男女ともに最も自信がある。





[39] 注 38 と同じ理由で、4 か国の出力。

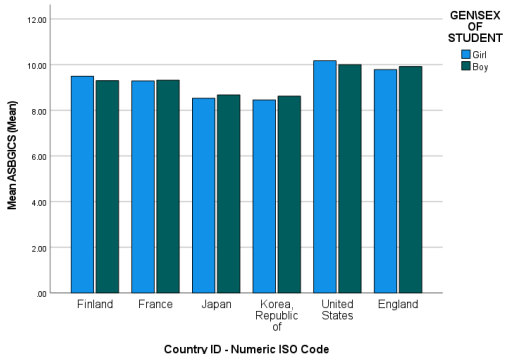


図 2-5：TIMSS_小 4 理科_「指導の明確さ」尺度（ASBGICS）

・フィンランドと米国は、女子の方が男子より、理科の授業が明確と感じている。両国は女子の方が男子より理科に自信があり、米国は女子が男子より理科を好む。

・日本と韓国の棒が低く、両国の理科の授業は明確さが相対的に乏しい。

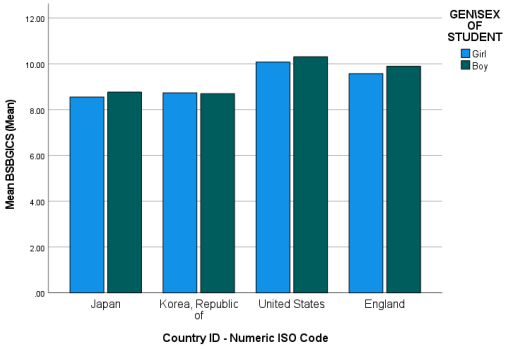


図 2-6：TIMSS_中 2 理科_「指導の明確さ」尺度（BSBGICS）[40]

・韓国のみ、女子の方が男子より、理科の授業が明確と感じている。

【算数・数学の分析結果】

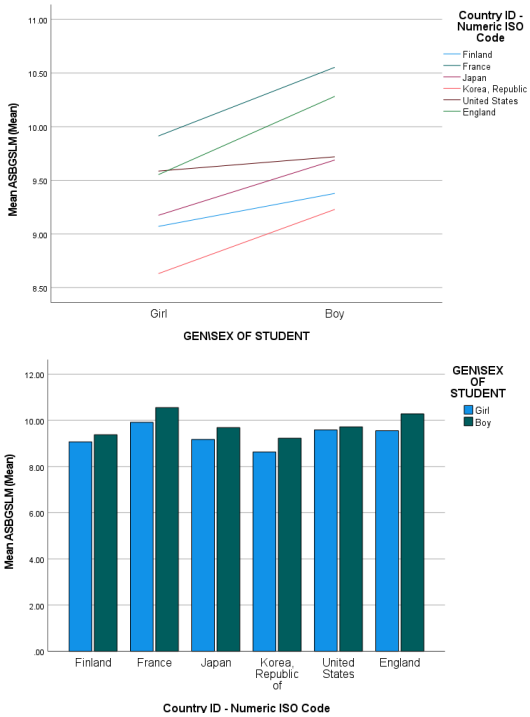


図 2-7：TIMSS_小 4 算数_「好き」尺度（ASBGSLM）

・6か国とも、男子の方が女子より、算数が好き。

・韓国と日本は、算数を好きな傾向が、他国よ

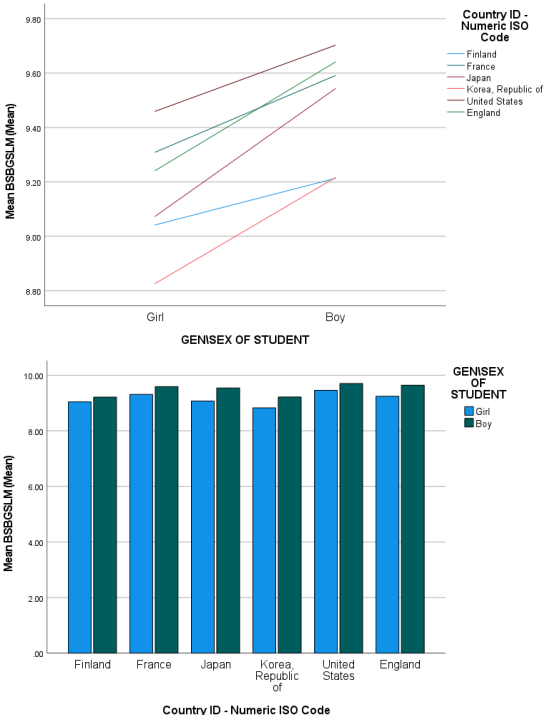


図 2-8：TIMSS_中 2 数学_「好き」尺度（BSBGSLM）

・6か国とも、男子の方が女子より、数学が好き。

[40] 注 38 と同じ理由で、4 か国の出力。

り低い。	

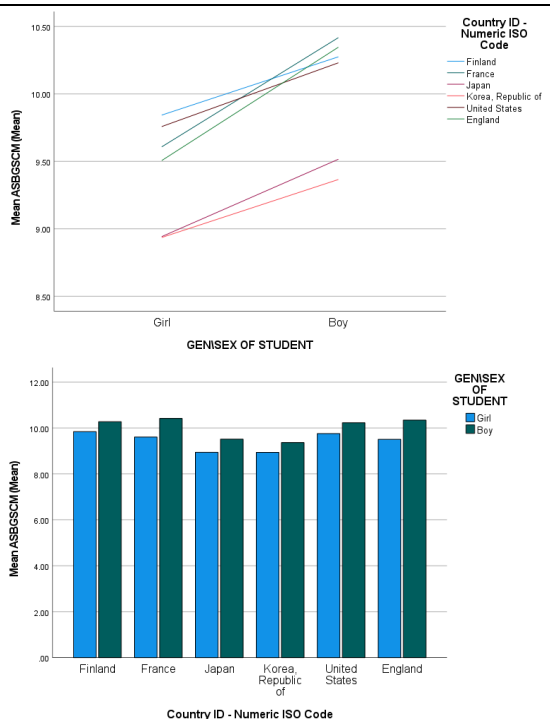


図 2-9：TIMSS_小 4 算数_「自信」尺度（ASBGSCM）

・6か国とも、男子の方が女子より、算数に自信がある。

・日本と韓国は、他国より算数への自信が低い（TIMSSスコアは高いにもかかわらず）。

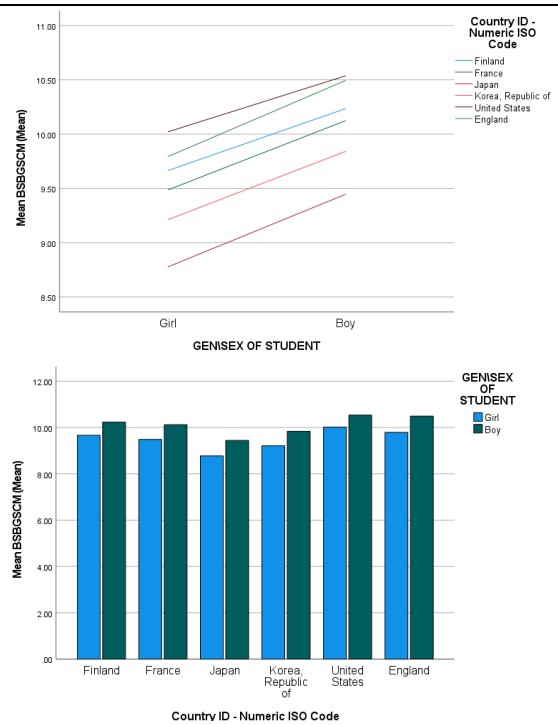
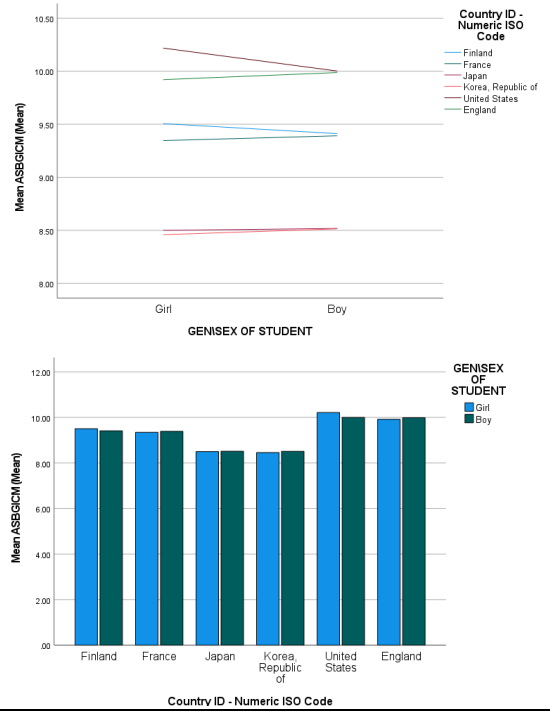


図 2-10：TIMSS_中 2 数学_「自信」尺度（BSBGSCM）

・6か国とも、男子の方が女子より、数学に自信がある。

・日本と韓国は、他国より数学への自信が低い（TIMSSスコアは高いにもかかわらず）。



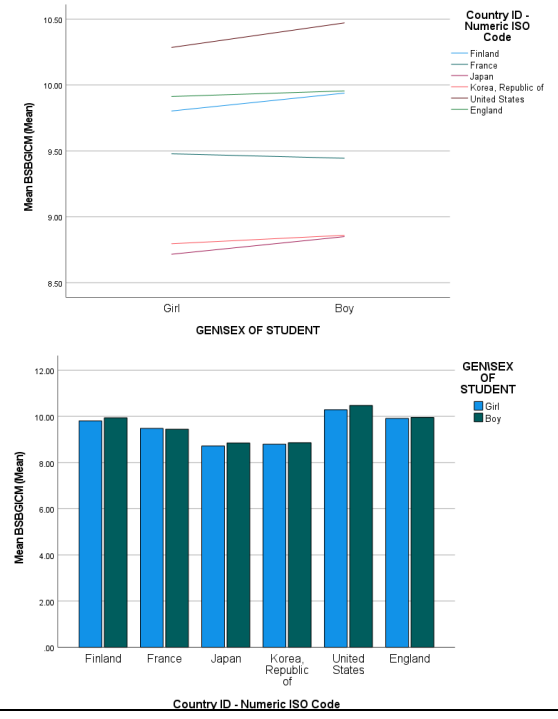


図 2-11：TIMSS_小 4 算数_「指導の明確さ」尺度（ASBGICM）	図 2-12：TIMSS_中 2 数学_「指導の明確さ」尺度（BSBGICM）
・フィンランドと米国は、女子の方が男子より、算数の授業が明確と感じている。	・フランスのみ、女子の方が男子より、数学の授業が明確と感じている。

【TIMSS2019_中 2 の結果】

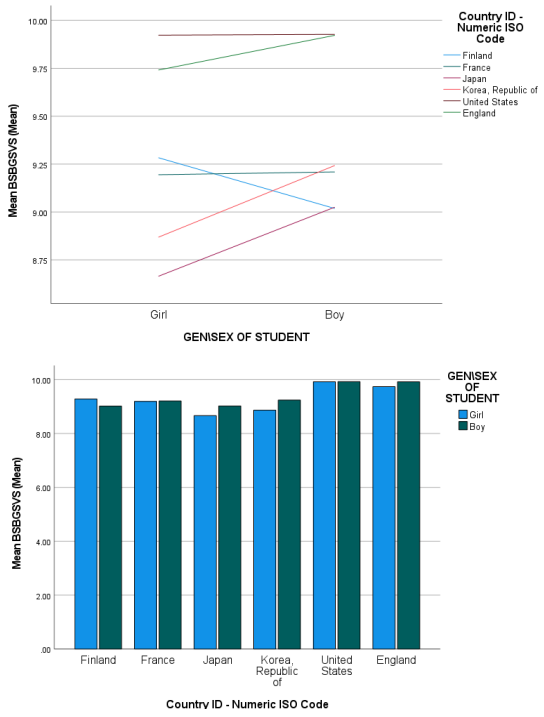


図 2-13：TIMSS_中 2 理科_「価値」尺度（BSBGSVS）

・フィンランドは、女子の方が男子より、理科に価値を認める傾向が高い。

・日本と韓国は、米国や英国より、男女ともに理科に価値を認める傾向が低い。

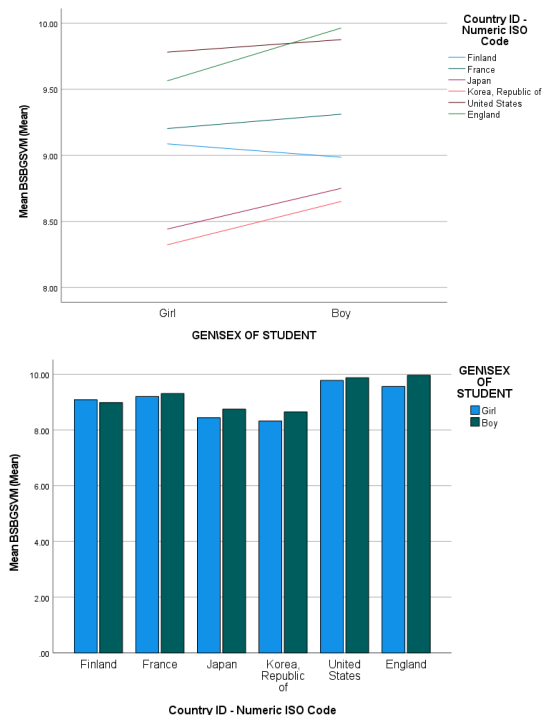


図 2-14：TIMSS_中 2 数学_「価値」尺度（BSBGSVM）

・フィンランドは、女子の方が男子より、数学に価値を認める傾向が高い。

・日本と韓国は、米国や英国より、男女ともに数学に価値を認める傾向が低い。

各尺度を構成する質問項目

※回答選択肢は、「強くそう思う」「そう思う」「そう思わない」「まったくそう思わない」の 4 件法である。

小 4 理科「好き」尺度	ア）理科の勉強は楽しい
	イ）理科の勉強をしなくてもよければいいのにと思う
	ウ）理科はたいくつだ
	エ）理科でおもしろいことをたくさん勉強している
	オ）わたしは、理科がすきだ
	カ）理科の授業が楽しみだ
	キ）理科はわたしに世の中の仕組みを教えてくれる
	ク）理科の実験をするのがすきだ
	ケ）理科はわたしのすきな教科の一つだ

※中 2 理科では、仮名と漢字の表記が異なるだけである。

小 4 算数「好き」尺度	ア）算数の勉強は楽しい
	イ）算数の勉強をしなくてもよければいいのにと思う
	ウ）算数はたいくつだ
	エ）算数でおもしろいことをたくさん勉強している
	オ）わたしは、算数がすきだ
	カ）わたしは数字に関する学校の勉強はどれもすきだ
	キ）わたしは算数の問題をとくのがすきだ
	ク）算数の授業が楽しみだ
	ケ）算数はわたしのすきな教科の一つだ

※小 4 理科「好き」尺度と質問内容が異なるのは、カ）キ）である。
※中 2 数学では、教科名が「数学」になり、仮名と漢字の表記が異なるだけである。

小 4 理科「自信」尺度	ア）理科の成績はいつもよい
	イ）わたしは、クラスの友だちよりも理科をむずかしいと感じる
	ウ）わたしは理科が苦手だ
	エ）理科でならうことはすぐにわかる
	オ）先生はわたしに理科がよくできると言ってくれる
	カ）わたしには、理科はほかの教科よりもむずかしい
	キ）理科はわたしをこまらせる

※中 2 理科では、ウ）が「理科は私の得意な教科ではない」になり、「私は理科の難しい問題を解くのが得意だ」という項目が加わっている。
※上記以外は、仮名と漢字の表記が異なるだけである。
※小 4 算数「自信」尺度は、小 4 理科「自信」尺度の教科名を変え、「算数はわたしをイライラさせる」「わたしは算数のむずかしい問題をとくのが得意だ」の 2 項目が加わっている。
※中 2 数学「自信」尺度は、小 4「自信」尺度と同じ質問項目で、仮名と漢字の表記が異なるだけである。

小 4 理科「指導の明確さ」尺度	ア）先生がわたしに何を期待しているかわかっている
	イ）わたしの先生はわかりやすい
	ウ）先生は私の質問にはっきりした答えを返してくれる
	エ）先生は理科の説明がうまい
	オ）先生は、わたしたちが学習するのを助けるためにいろいろなことをしてくれる
	カ）先生は、わたしたちがわからなかったときにもう一度せつめいしてくれる

※中 2 理科では、「先生は、新しい授業ですでに私が知っていることと結びつけてくれる」という 1 項目が加わっている。それ以外は、仮名と漢字の表記が異なるだけである。
※小 4 算数では、教科名が異なるだけである。
※中 2 数学は、中 2 理科と教科名が異なるだけである。

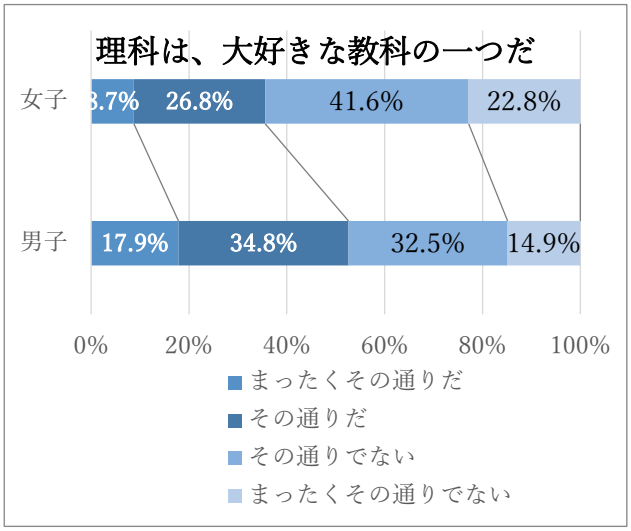
中 2 理科	a）理科を勉強すると、日常生活に役立つ

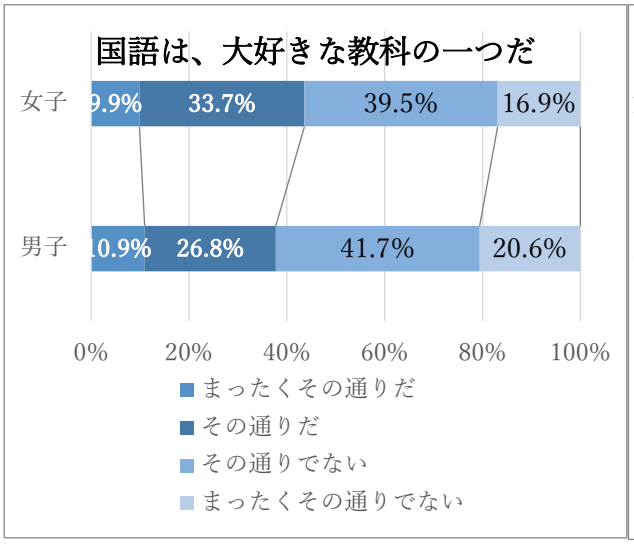
「価値」尺度	b）他教科を勉強するために理科が必要だ
	c）自分が行きたい大学に入るために理科で良い成績をとる必要がある
	d）将来、自分が望む仕事につくために、理科で良い成績をとる必要がある
	e）理科を使うことが含まれる職業につきたい
	f）世の中で成功するためには理科について勉強することが重要である
	g）理科を勉強することで、大人になってより多くの職業の機会が得られる
	h）私の両親は、私が理科で良い成績をとることが重要であると思っている
	i）理科の成績が良いことは大切だ

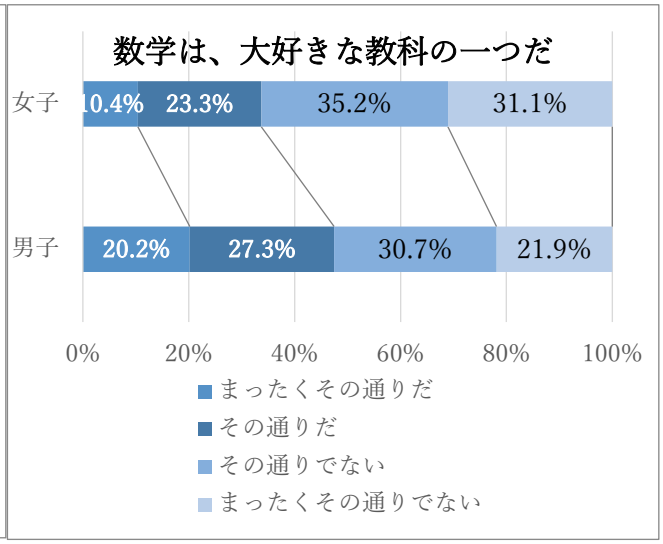
※中２数学では、教科名が異なるだけである。

４．５．理科・国語・数学の学習に対する興味・関心の男女比較 —国際調査 PISA2022 での検討—

PISA2022 の生徒質問調査の「ST268　数学・国語・理科の中での、好きな教科と自己認知（願望）」のうち以下の項目の回答結果を男女比較。







任意に抽出したフィンランド、フランス、日本、韓国、米国、英国のうち、フィンランド以外の国では、日本と同様、理科・数学が好きと回答する割合は女子の方が小さく、国語が好きと回答する割合は女子の方が大きいという傾向が確認された。

「●●は、大好きな教科の１つだ」という項目への回答状況（単位は%）

フィンランド		まったくその通りだ	その通りだ	その通りでない	まったくその通りでない
理科	女子	9.6	30.5	41.2	18.7
	男子	8.6	31.9	40.9	18.6
国語	女子	6.4	23.5	47.9	22.2
	男子	3.1	18.3	50.7	27.9
数学	女子	8.4	20.4	36.7	34.6
	男子	11.4	26.9	35.1	26.6

フランス		まったくその通りだ	その通りだ	その通りでない	まったくその通りでない
理科	女子	13.7	26.5	31.1	28.6
	男子	19.3	35.4	26.1	19.1
国語	女子	11.2	30.8	36.9	21.2
	男子	6.1	21.6	41.6	30.7
数学	女子	13.4	21.9	24.4	40.3
	男子	21.0	28.3	24.1	26.5

日本		まったくその通りだ	その通りだ	その通りでない	まったくその通りでない
理科	女子	8.7	26.8	41.6	22.8
	男子	17.9	34.8	32.5	14.9
国語	女子	9.9	33.7	39.5	16.9
	男子	10.9	26.8	41.7	20.6
数学	女子	10.4	23.3	35.2	31.1
	男子	20.2	27.3	30.7	21.9

韓国		まったくその通りだ	その通りだ	その通りでない	まったくその通りでない
理科	女子	13.8	31.0	34.9	20.4
	男子	19.1	33.8	29.3	17.8
国語	女子	11.7	37.7	36.6	13.9
	男子	10.0	34.2	38.1	17.7
数学	女子	9.0	23.8	35.4	31.7
	男子	15.9	29.3	28.5	26.3

米国		まったくその通りだ	その通りだ	その通りでない	まったくその通りでない
理科	女子	14.1	34.7	36.1	15.0
	男子	16.2	42.1	29.6	12.1
国語	女子	14.3	39.1	33.4	13.2
	男子	7.4	31.0	41.4	20.2
数学	女子	11.6	27.9	31.3	29.1
	男子	14.5	32.0	30.0	23.5

英国		まったくその通りだ	その通りだ	その通りでない	まったくその通りでない
理科	女子	12.1	27.8	37.1	23.0
	男子	13.7	35.5	34.6	16.2
国語	女子	14.8	34.6	34.8	15.7
	男子	6.7	29.7	42.6	20.9
数学	女子	10.1	25.1	38.1	26.7
	男子	15.3	35.6	32.9	16.2

５．理科調査の品質検証

５．１．平成 24、27、30、令和 4 年度の理科調査の品質検証

リサーチクエスチョン：

平成 24、27、30、令和 4 年度の全国学力・学習状況調査の小 6 と中 3 の理科調査のテスト精度は、一定の品質を担保しているか？

※　全国学力・学習状況調査の結果分析が意義あるものになるためには、根拠となる調査問題（広義の「テスト」）に一定の精度がなくてはならない。

分析結果：

平成 30 年度小 6 と令和 4 年度中 3 の理科の調査問題は、信頼性において若干の課題があった。他の年度では一定の信頼性が確保されていた。

平成 30 年度小 6 と令和 4 年度中 3 の理科の調査問題は、他の年度の調査問題と比較して、信頼性係数とテスト情報量が小さい[41]。国が実施する悉皆調査としては、理科調査では、信頼性係数 $\alpha \geqq 0.8$ を確保するのが望ましい[42]。

【検証 01】信頼性係数（クロンバックのα）による検証

EasyEstimation が分析出力する信頼性係数（クロンバックのα）は無回答を誤答に含むため、無回答を含めない場合よりも若干高めに出る[43]。無回答を含む場合の理科の基準値を $\alpha \geqq 0.80$ とする。

信頼性係数は問題数の影響を受けるため、まず理科調査の問題数を確認する。

表 5-1：理科調査の問題数一覧

	H24	H27	H30	R4
EL	24	24	16	17
JH	26	25	27	21

単位：問

EasyEstimation が出力した各年度理科の信頼性係数は表 5-2 である。

表 5-2：信頼性係数（α）一覧

	H24	H27	H30	R4
EL	0.82	0.83	0.73	0.82
JH	0.85	0.88	0.85	0.77

[41] 「信頼性係数（クロンバックのα）」と「テスト情報量」については、本章末の用語解説 B を参照されたい。

[42] 信頼性係数 $\alpha > 0.8$ の基準値についても、用語解説 B を参照されたい。

[43] Cf.，田端，2022，p. 36.

【検証 01】の結果と評価

表 5-2 で、薄い赤でハイライトした 2 か所（平成 30 年小 6、令和 4 年中 3）で基準値を下回っている。令和 4 年度小 6 は全 17 問でありながら、α=0.82 となり基準値を上回っており、調査問題の品質を評価できる。全体としては、平成 30 年度小 6 理科の調査にやや課題があるものの、信頼性係数の観点からは、一定の精度があると評価できる。ただし、全 8 回の調査項目のうち 2 回で基準値を下回っており、この結果を参考に、今後基準値を下回らないことが望ましい。

【検証 02】テスト情報量による検証

EasyEstimation は IRT 分析ソフトであり、各項目の困難度の高低を分析できる。このソフトの「項目特性曲線／テスト情報曲線」の分析では、「THETA（学力θ）」と「TestInformation（テスト情報量）」が出力される。そこで、出力されたテスト情報量を、統計ソフト R にてテスト情報量曲線として可視化してみる[44]。

※ 全国学力・学習状況調査の教科に関する調査は、いわゆる「テスト」ではないが、調査項目の情報量曲線は一般に「テスト情報量曲線」と呼ばれるため、この名称を使った。

【検証 02】の結果と評価

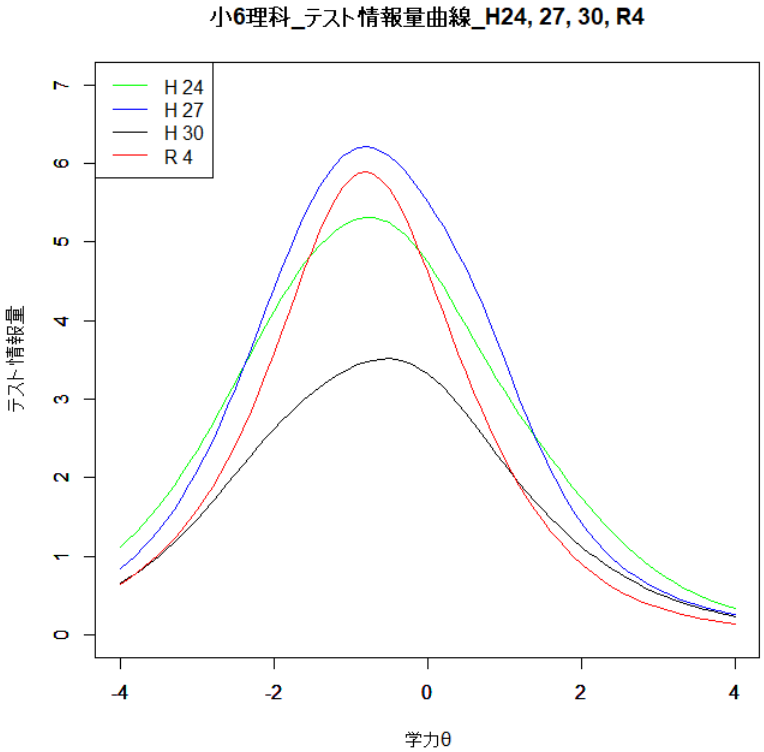


図 5-1：小 6 理科テスト情報曲線の比較

図 5-1 は、小 6 理科の 4 か年のテスト情報量曲線である。平成 30 年度（黒）の曲線の山が最も低

[44] 「テスト情報量曲線」の意味とグラフの読み方については、用語解説 B を参照されたい。

く、4 か年のうち最もテスト情報量が少ないことがわかる。表 5－2 のα係数も、平成 30 年度は小 6 理科 4 か年のうち最も低い 0.73 である。平成 30 年度小 6 理科は、α係数でもテスト情報量でも、信頼性が低いことになる。

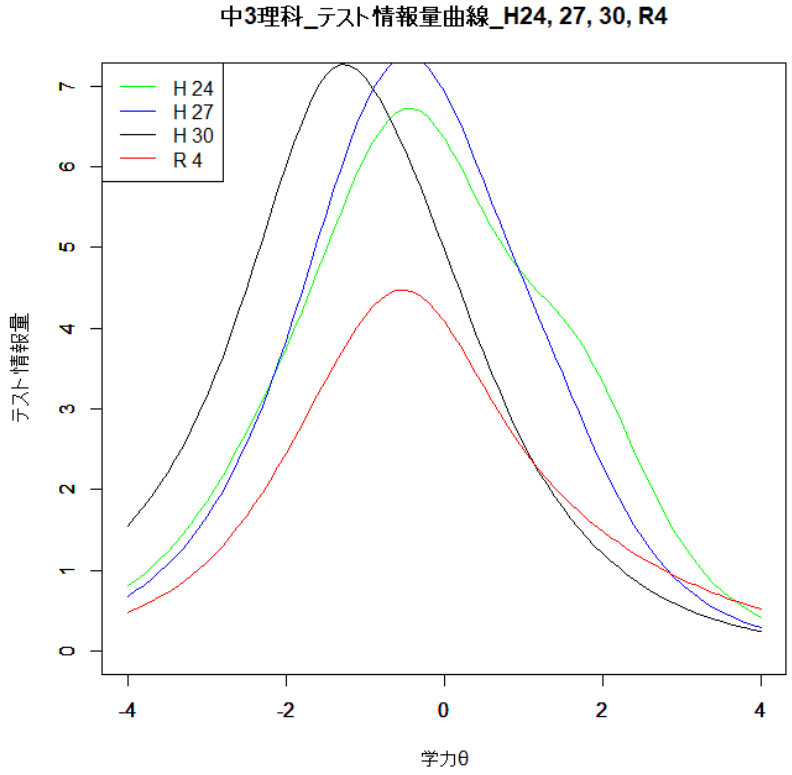


図 5-2：中 3 理科テスト情報曲線の比較

図 5-2 は、中 3 理科の 4 か年のテスト情報量曲線である。令和 4 年度（赤）の曲線の山が最も低く、4 か年のうち最もテスト情報量が少ないことがわかる。表 5－2 のα係数も、令和 4 年度は、中 3 理科 4 か年のうち最も低い 0.77 である。中 3 理科では、令和 4 年度が、α係数でもテスト情報量でも、信頼性が低いことになる。

５．２．令和 4 年度、3 教科の調査問題の品質比較

リサーチクエスチョン：

全国学力・学習状況調査の理科の調査問題は、国語や算数・数学と比較して、どの程度の信頼性をもつか。令和 4 年度で 3 教科の信頼性とテスト情報量を比較する。

分析結果：

信頼性係数とテスト情報量の点から、令和 4 年度中 3 理科には、改善の余地がある。令和 4 年度小 6 理科と同程度の信頼性とテスト情報量を確保するのが望ましい。

分析方法：

令和 4 年度の小 6 と中 3 の理科と国語と算数・数学のデータを EasyEstimation にかけ、信頼性係数とテスト情報量を計算・可視化する。

【検証 01】EasyEstimation で出力される信頼性係数（クロンバックのα）による、令和 4 年度の小 6 中 3 各教科の信頼性比較

基準値としては、国語の調査問題は算数・数学より信頼性を確保するのが難しいことから、国語の基準値は 0.75 とする[45]。算数・数学は、理科と同じ 0.80 とする。

表 5-3：令和 4 年度、3 教科の信頼性係数（α）

	理科	国語	算数・数学
R4_小 6	0.82	0.80	0.82
R4_中 3	0.77	0.77	0.83

評価としては、令和 4 年度は 3 教科中、中 3 理科の調査問題のみ、信頼性の基準値に届いていない。この他はすべて、基準値を超えている。

【検証 02】テスト情報量の比較

EasyEstimation が出力するテスト情報量を、統計ソフト R で可視化。

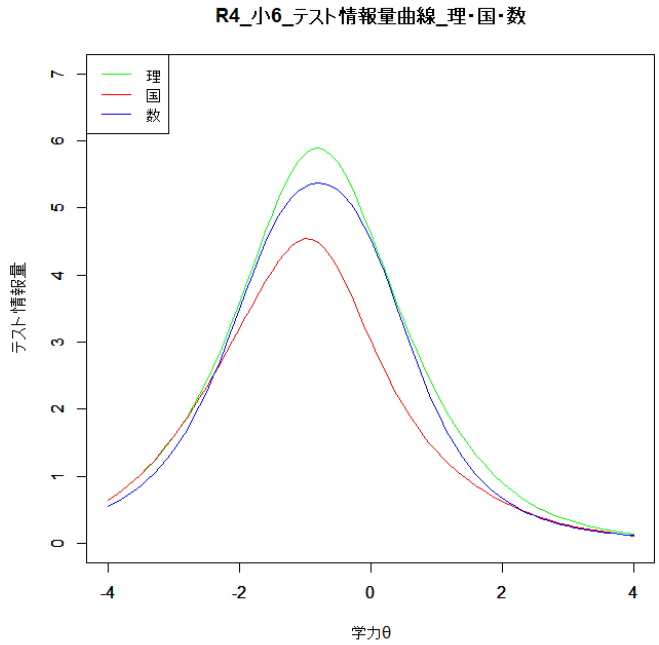


図 5-3：令和 4 年度小 6、3 教科のテスト情報量曲線比較

[45] Cf., 田端, 2022, p. 34. 用語解説 B も参照されたい。

図 5-3 は、令和 4 年度小 6 の 3 教科のテスト情報量曲線である。国語（赤）のテスト情報量が最も少ない。理科は 3 教科の中で、テスト情報量の最大値が最も大きい。α係数も 0.82 と高く、テスト情報量と信頼性係数の二つの観点から、全 17 問でありながら信頼性の高い問題であったと評価できる。

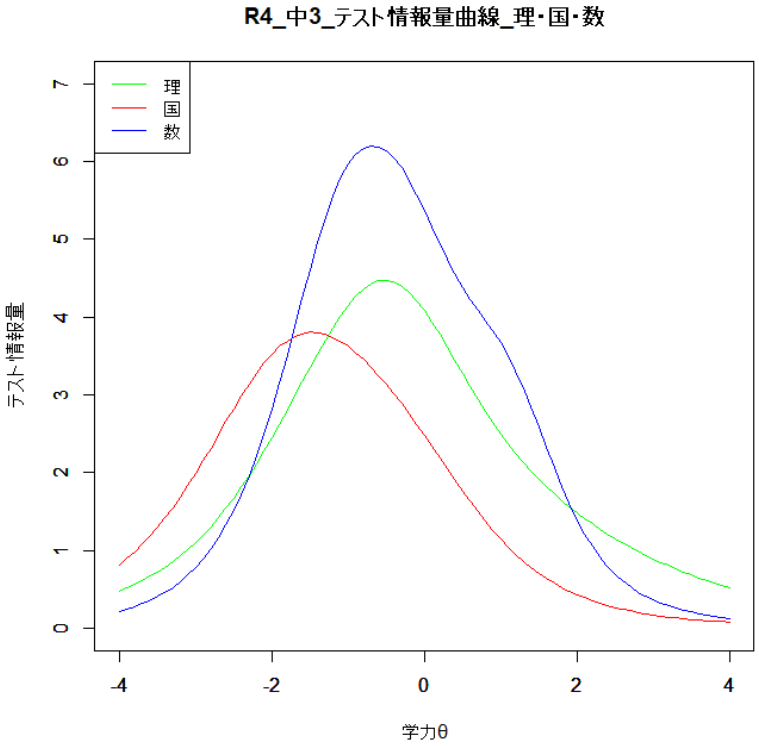


図 5-4：令和 4 年度中 3、3 教科のテスト情報量曲線比較

図 5-4 は、令和 4 年度中 3 の 3 教科のテスト情報量曲線である。テスト情報量の最大値は、国語（赤：3.80）が最も小さく、次に理科（緑：4.48）が小さい。表 5-3 より、α係数は、国語が 0.77、理科も 0.77 である。理科は、α係数で基準値を下回っており、テスト情報量も大きくないことから、質改善が望ましい。国語の基準値は 0.75 としているため、信頼性係数では基準値は超えているが、テスト情報量の少なさから、改善の余地がある結果である。

５．３．令和 4 年度理科、中 3 と小 6 の項目特性曲線の比較

リサーチクエスチョン：

令和 4 年度理科の教科調査で、小 6 は 17 問、中 3 は 21 問と中 3 の問題数が 4 問多いにもかかわらず、信頼性係数αは小 6 が 0.82、中 3 は 0.77 と小 6 の方が高い。令和 4 年度中 3 理科調査は、問題数が多いにもかかわらず、どうして同年度小 6 理科調査の信頼性係数より低いのか。

分析結果：

令和 4 年度中 3 理科の問題項目には、難易度が高すぎる問題と識別力が低い問題が多く、結果として調査問題冊子の信頼性係数の低下につながっている。

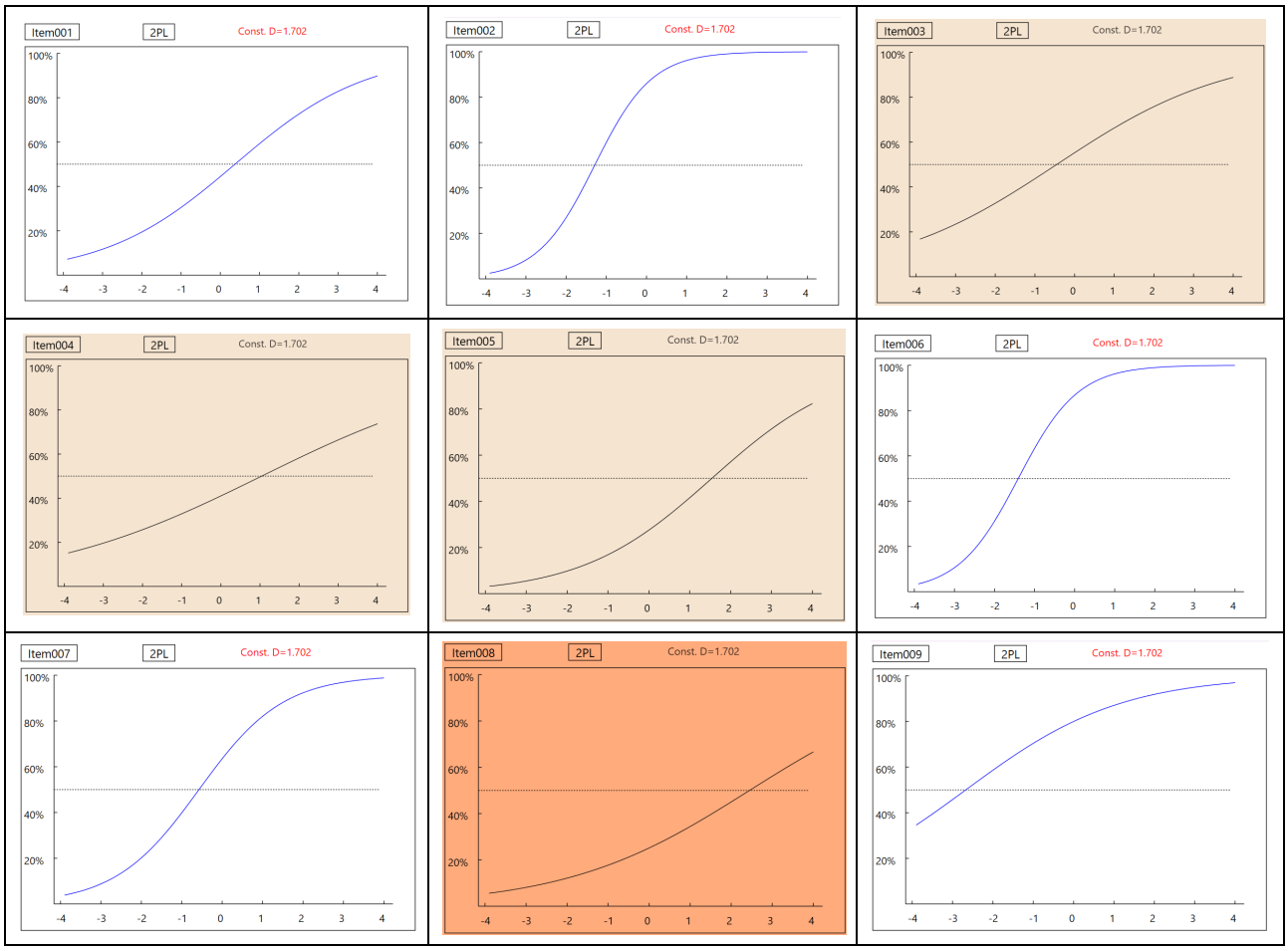
分析方法：

EasyEstimation で出力される項目特性曲線[46]を、令和 4 年度の小 6 と中 3 で比較する。

【検証】

令和 4 年度理科中 3 の各問題項目の項目特性曲線は、図 5-5 になる。EasyEstimation の出力グラフを転記し一覧化した。

難易度が極端に高い（平均より 2 シグマ以上高い）項目を濃い赤で、識別力が特に低い項目を薄い赤でハイライトした。



[46] 項目特性曲線については、用語解説 B を参照されたい。

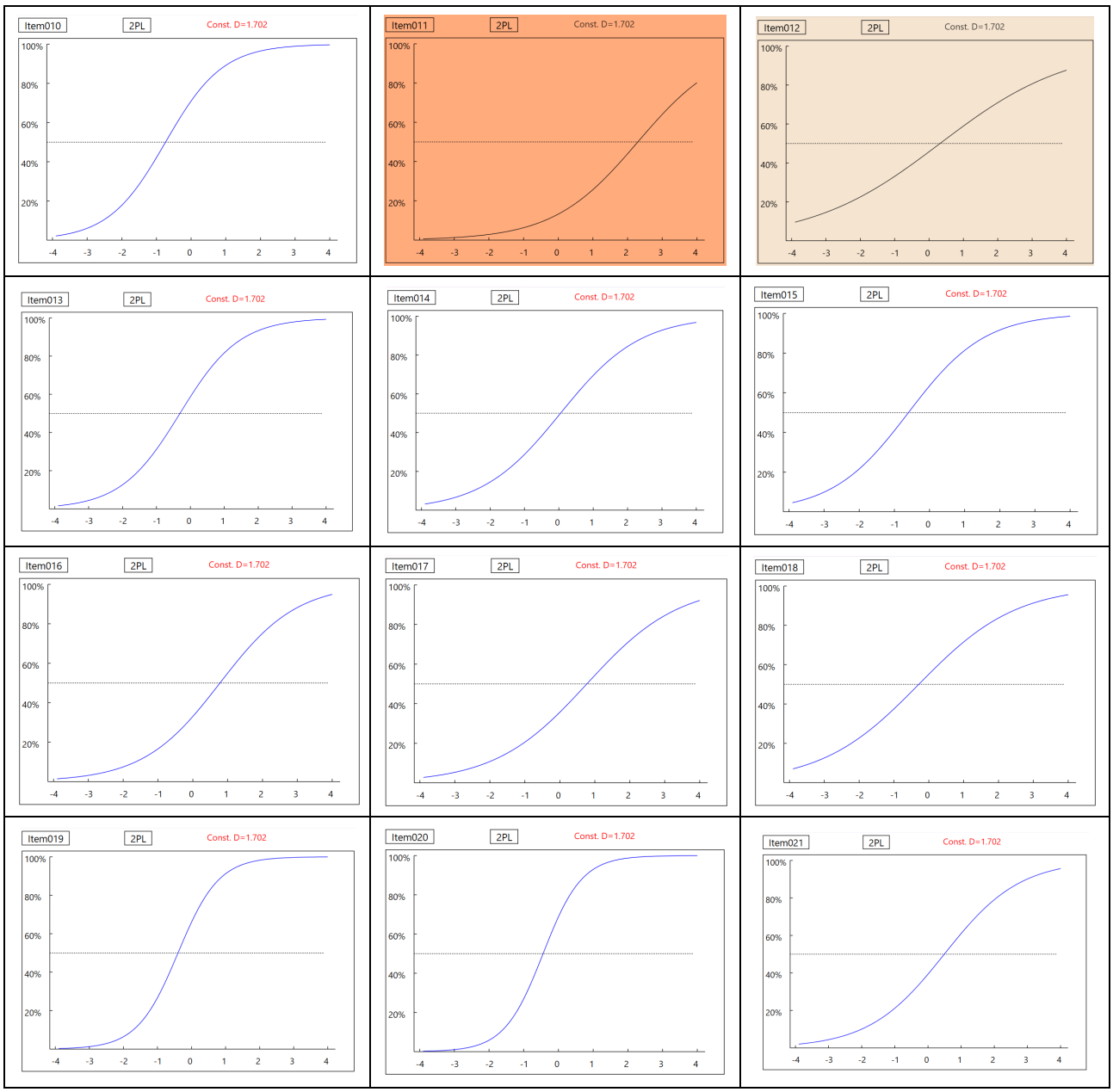


図 5-5：令和 4 年度中 3 理科の項目特性曲線一覧

Item008（問題番号３（３））と Item011（問題番号５（１））は、平均より 2 シグマ以上高く、偏差値にして 70 以上の難易度になる。標準正規分布の割合によれば、平均より 2 シグマ以上にはおよそ 2%しか入らない。この 2%を識別する項目は、全 21 問の冊子デザインからすると、相応しいとはみなしがたい。

これら 2 項目を、「令和 4 年度　全国学力学習状況調査報告書　中学校理科」で確認すると、正答率 24.9%と 15.5%である（同報告書，p.11）[47]。正答率は、問題項目の難易度を推測する 1 つの指標である。しかし、2PLM で特定した難易度からすると、この正答率には、あてずっぽうの正解が少なからず含まれていると推測できる。これら 2 項目の識別力も、この推測を支持する。

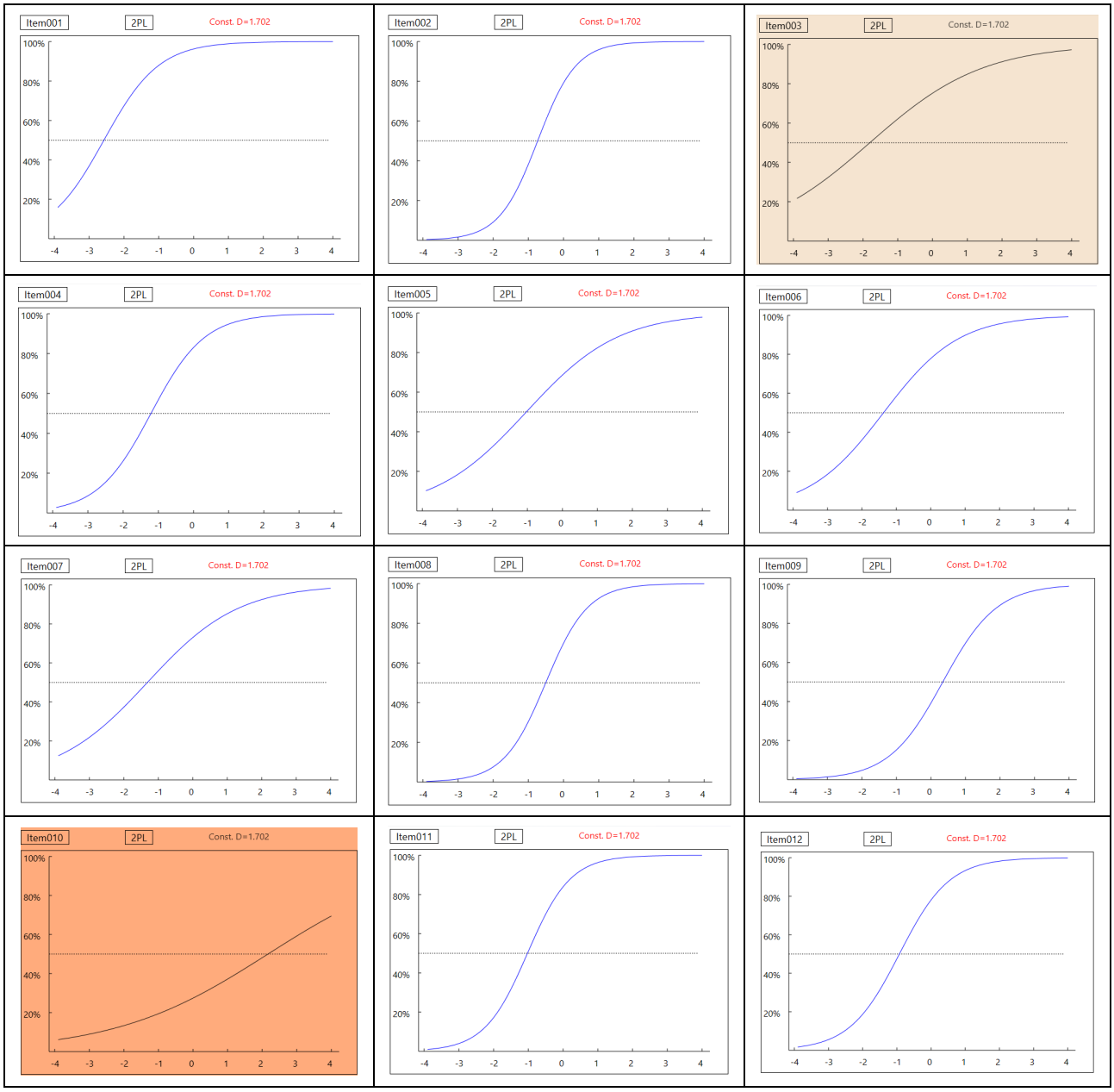
以上から、この 2 項目には実質的な識別力がないと評価せざるを得ない。

[47] https://www.nier.go.jp/22chousakekkahoukoku/report/data/22msci.pdf ［2024.03.11 最終閲覧］

薄い赤でハイライトした Item003（問題番号２（１））、Item004（問題番号２（２））、Item005（問題番号２（３））、Item012（問題番号５（２））の５項目は、曲線の傾きが特になだらかで、識別力としては 0.5 を切るほどの低さである。Item001 や Item017 や Item018 もほぼ同様に傾斜がなだらかで、識別力が弱い。

ハイライトした７項目を差し引くと、識別力が一定以上の項目は 15 問となる。

対して、令和４年度小６理科の項目特性曲線一覧は、図 5-6 になる。



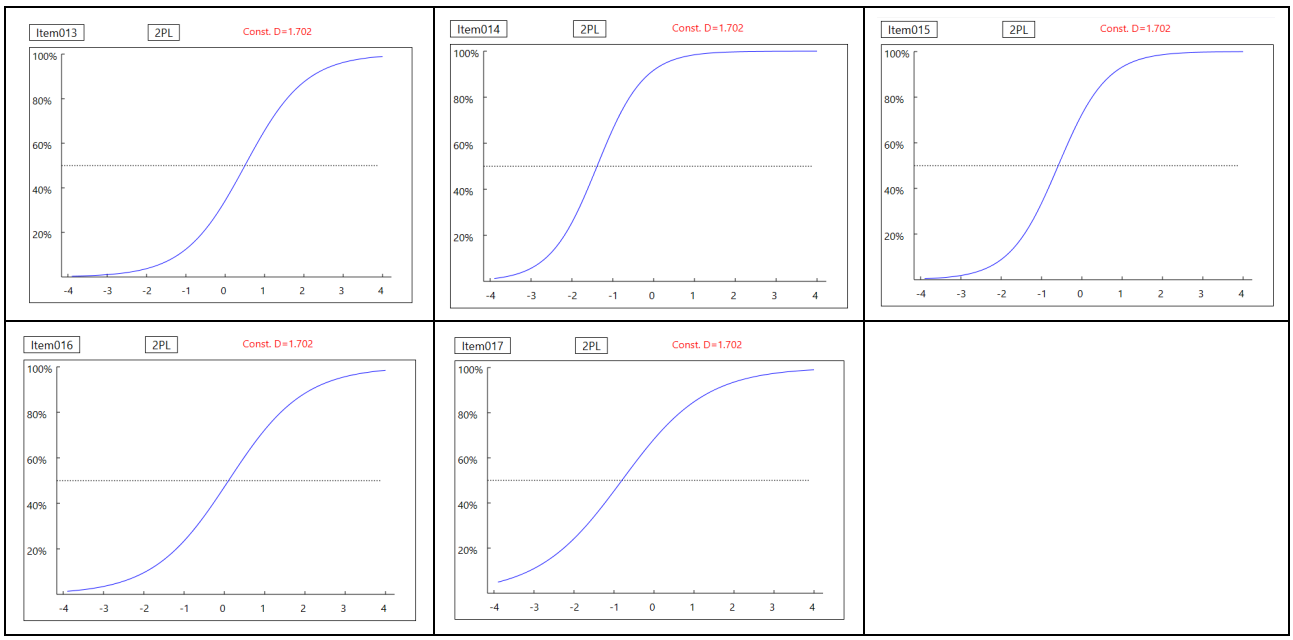


図 5-6：令和 4 年度小 6 理科の項目特性曲線一覧

全体として、中 3 理科よりも、曲線が立ち上がっている傾向がうかがわれる。図 5−5 と同じ基準でハイライトした。

難易度が極端に高い項目は Item010（問題番号３（１））、識別力が特に低い項目は Item003（問題番号１（３））の 2 項目である。全項目数からこれら 2 項目を差し引くと、一定の識別力があるのは 15 問になる。

令和 4 年度理科調査の中 3 と小 6 を、難易度と識別力の観点から比較すると、前者の項目数が多いにもかかわらず、実用的な難易度と識別力をもつ項目は、中 3 小 6 とも 15 問（項目）となり、信頼性係数の逆転が生じても納得できる結果である。

○用語解説 B

B-1 信頼性係数（クロンバックのα）

テストの「信頼性（reliability)」とは、「そのテストが測りたいものを本当に精度よく測定しているかどうか」[48]に関わる問題である。信頼性の測定には複数の方法があるが、「内的整合性」の測度は、「教育・心理統計」で「一番重要な信頼性の指標」とされる[49]。これは、「一度のテストの結果からテスト内の相関情報を利用して信頼性係数の推定値を求める方法」[50]である。その中でも代表的なものがクロンバック（Cronbach）のαであり、次の式[51]：

$$\alpha = \left(\frac{m}{m-1}\right)\left(1 - \frac{\sum_{j=1}^{m}\sigma_j^2}{\sigma_x^2}\right)$$

で求められる。mは項目数、σ_x^2はテスト得点の分散、σ_j^2は各項目の分散を示している。

αは、0から1までの値をとり、1に近いほど信頼性が高い。

右辺の

$$\left(\frac{m}{m-1}\right)$$

から、項目数（m）が多いほど、αが1に近づくことが分かる。また式右辺の

$$\left(1 - \frac{\sum_{j=1}^{m}\sigma_j^2}{\sigma_x^2}\right)$$

から、分母のテスト得点の分散（σ_x^2）が大きく、分子の各項目の分散の和（$\sum_{j=1}^{m}\sigma_j^2$）が小さいほど、αは1に近づくこともわかる。

0から1までのα係数をどう評価するかは、「基準値」の問題である。α係数の基準値にも種々の議論がある。光永悠彦氏は、「一般的な心理テストの場合」は0.8以上、「試験問題冊子の信頼性」はおおむね0.9から0.95が望ましく、「研究目的」では0.7から0.8でもよい、と要約している[52]。また、柴山直氏は、「たとえばアメリカでは採用試験や選抜試験は0.9以上の信頼性をもつよう法律で求めている」と指摘し、「学力検査なら0.8以上が必要となるが、パーソナリティ検査なら0.7以上でもまあ仕方ないであろう」としている[53]。α係数の基準値を広範に検討したTaber(2018)は、「excellent(0.93-0.94)」「strong (0.91-0.93)」「reliable (0.84-0.90)」「robust (0.81)」「relatively high (0.70-0.77)」「slightly low (0.68)」「not satisfactory (0.4-0.55)」とまとめている[54]。「頑健（robust)」以上と評価するには、0.80以上必要である。0.70を下回ると、「若干低い（slightly low)」となる。0.70-0.77は微妙で、「比較的高い（relatively high)」というのがTaberのまとめである。0.70以上あればよいという評価も成り立たないわけではない。

しかし、国が毎年実施する悉皆調査で、学校や各自治体ひいては社会全体への影響力が大きな調査であることからすれば、加えて無回答を誤答に含む計算であり、α係数が高めに出る統計手法であることからすれば、理数教科で0.80、国語で0.75は、無難な基準値であり、それほど高いハードルではない。

---

[48] 柴山直（n.d.），講義メモ　古典的テスト理論，p.8.
[49] 柴山（n.d.），p.14.
[50] 柴山（n.d.），p.14.　「テストない」を「テスト内」に改めて引用した。
[51] 柴山（n.d.），p.14.
[52] 光永（2018），p.92.
[53] 柴山（n.d.），p.16.
[54] Taber, K.S. (2018), The Use of Croncach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, 48, Springer, p.1278.

B-2 テスト情報量とテスト情報量曲線

「テスト情報量(test information)」[55]とは、ある問題冊子が受検者のどの学力（θ）に対してどの程度の情報をもたらすかを示す指標である。横軸に学力値、縦軸にテスト情報量をとり、ある問題冊子の情報量の多寡を分布曲線で表したものが「テスト情報量曲線（test information curve)」である。山が高いほど情報量が多く、学力の識別力が高い。テスト情報量曲線の見方としては、曲線の山のピークがどこにあるかにより、その問題冊子がどの学力レベルの識別に長けているかを確認し、山のピークの高さを比較することで、複数の問題冊子の識別力の高低を比較・判別できる。本調査研究の図 5-1「小6 理科テスト情報曲線の比較」を例にとれば、4 つの年度の問題冊子はいずれも、平均（0）よりも低いところに山のピークがあり、平均よりも低い学力値の識別に長けており、平成 27 年度（青）のテスト情報量が最も大きく、平成 30 年度（黒）が最も小さい。

なお、テスト情報曲線を活用した調査問題の「精度」検証は、令和 3 年度経年変化分析調査でも実施されている[56]。

B-3 項目特性曲線

「項目特性曲線（Item Characteristic Curve: ICC)」とは、項目反応理論において問題項目ごとに推定される学力値（横軸）と正答確率（縦軸）の関係を表わす曲線である。「ロジスティック曲線」「項目反応関数」とも呼ばれる。項目反応理論のモデルは幾つかあるが、EasyEstimation を活用した本調査研究は、2 パラメタ・ロジスティックモデル（「2PLM」と略記）を採用している。2PLM は、項目の識別力（a）と難易度（b）の 2 つの項目パラメタをもつ次の数式で表される[57]。D は定数で、1.702 である。

$$P(\theta|a,b) = \frac{1}{1+\exp\left(-Da(\theta-b)\right)}$$

識別力（a）が同じ 0.5 で、難易度（b）が異なる項目 01（難易度 b=-1.0）と項目 02（難易度 b=0.0）を作図すると図 5-7 になる。作図では、便宜的に D=1.7 とした。項目 02 の難易度は平均値（0.0）であり、項目 01 はそれより 1 シグマ（σ）分簡単な問題という設定である。

---

[55] Cf.，光永悠彦，2018『テストは何を測るのか―項目反応理論の考え方―』ナカニシヤ出版，pp.117-119.

[56] Cf.，文部科学省（2022），pp.63-65.

[57] 光永(2018)，p.113.

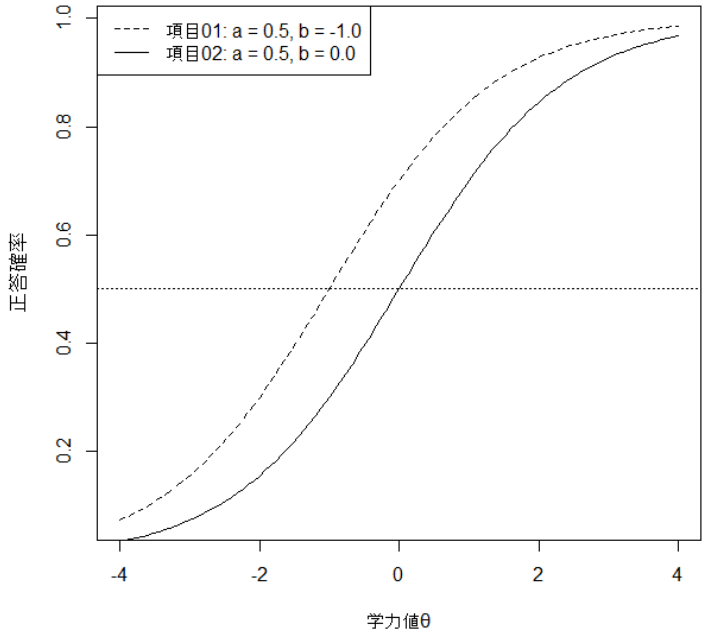


図 5-7：2PLM の項目特性曲線（同じ識別力で異なる難易度）

縦軸の 0.5 の目盛りの位置、つまり正答確率 50%の位置に横の点線を引いている。この横点線と曲線との交点が項目難易度になり、項目 01 の学力値 θ は-1.0、項目 02 の学力値 θ は 0.0 になっている。難易度が上がると、項目特性曲線は右に平行移動する[58]。

曲線の傾き具合は、識別力（a）を表わしている。同じ難易度 b=0.0 で、識別力（a）が異なる 3 つの項目を描画すると図 5-8 になる。項目 02 は a=0.5、項目 03 は a=1.0、項目 04 は a=1.5 であり、順に識別力を高くしている[59]。

[58] Cf.， 光永(2018)，p.114. Cf.， 文部科学省（2022），p.5.
[59] Cf.， 光永(2018)，p.115. Cf.， 文部科学省（2022），p.5.

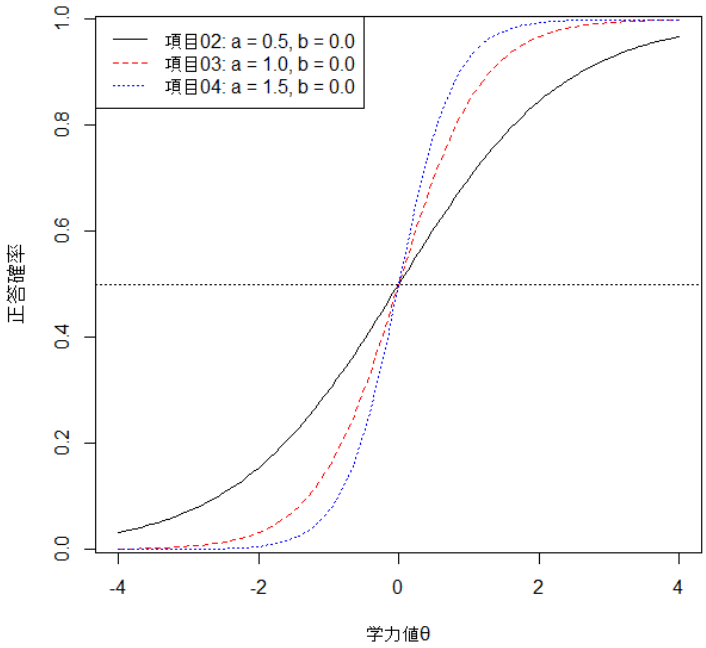


図 5-8：同じ難易度で異なる識別力の項目特性曲線

識別力が上がるにつれ、寝ていた曲線が立ち上がる（傾きが急になる）のがわかる。項目 04 では、学力値 θ が 0.0（平均）より 1 シグマ（偏差値にして 10）高い児童生徒（θ =1.0）の正答確率がほぼ 90%になり、1 シグマ低い児童生徒（θ =−1.0）の正答確率がほぼ 10%になる。つまり、学力値 θ =1.0 以上の児童生徒はほぼ全員が正答し、−1.0 以下の児童生徒はほぼ全員が誤答になる。これが識別力の高さである。対して項目 02 は、学力値 θ が平均より 1 シグマ高い児童生徒が約 70%しか正解せず（1 シグマ高い児童生徒でも約 30%が間違う）、逆に 1 シグマも低い約 30%の児童生徒も正答することになる。つまり、識別力の低い項目は、正答と誤答を分ける学力値を鋭く特定することができない。

６．生徒のウェルビーイングやいじめ反対意識の男女差国際比較

リサーチクエスチョン：

理科調査分析からは離れるが、生徒のウェルビーイングやいじめ反対意識に関して、男女差はあるのだろうか？それは国際比較するとどうなるか？

分析結果：

1. ウェルビーイング関係（図 6-1 から図 6-4 まで）は、可視化のグラフに結果の評価を記載。
2. いじめ反対意識に関連するすべての項目（図 5-5 から図 5-9 まで）につき、６か国すべてで、女子は男子より反対意識が高い。性差の可能性がある。

分析方法：

本調査研究２．４．および２．５．と同じ。

データは、ウェルビーイングやいじめ反対意識の質問紙調査がある PISA2018 を利用した。

結果の可視化と評価：

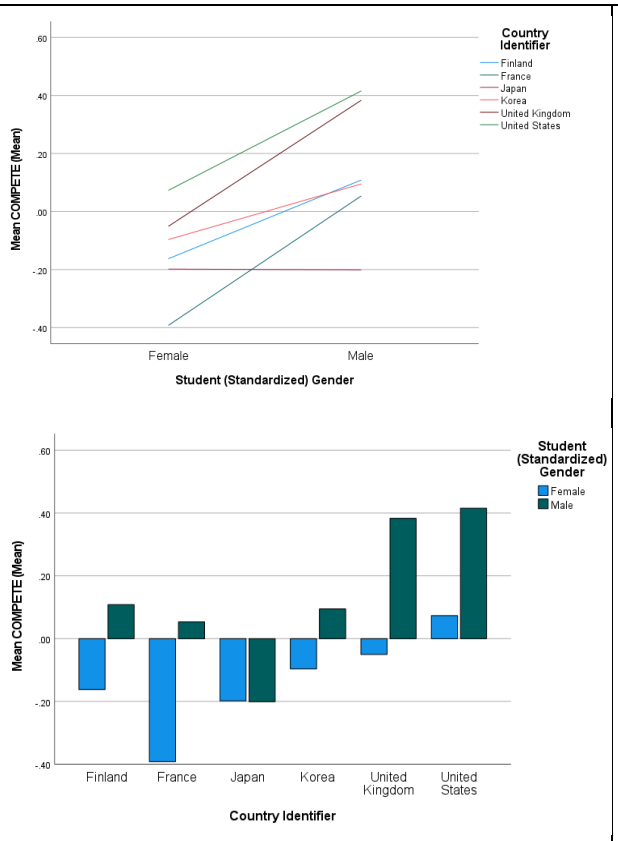


図 6-1：学校の競争的空気の度合い

1. 日本以外は男子が女子より圧倒的に競争にさらされている。
2. 日本は男女差なく、最も競争的でない学校風土である。

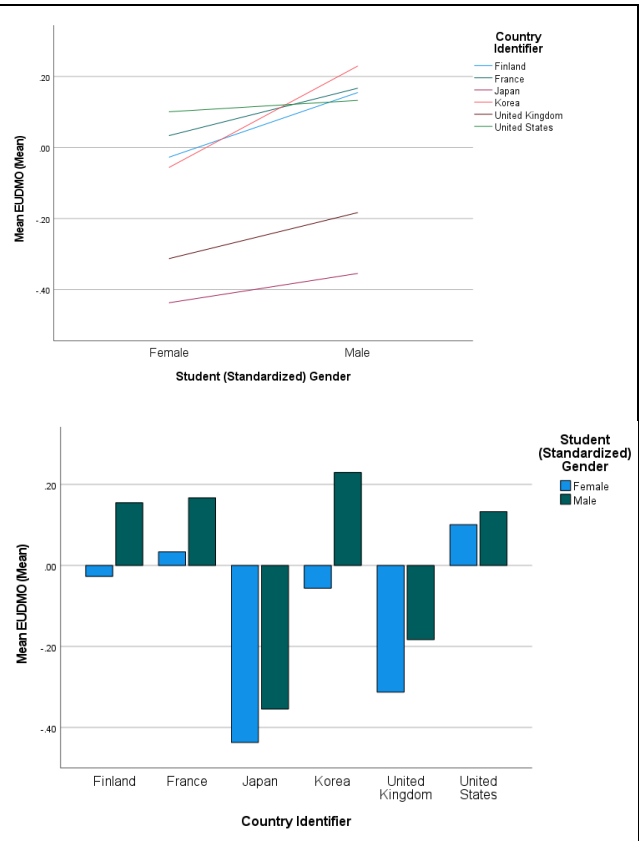


図 6-2：人生の意味

1. 日本の生徒が最も人生に意味を見いだせていない。
2. 6 か国いずれも、女子は男子より人生の意味を見いだしにくい傾向がある。

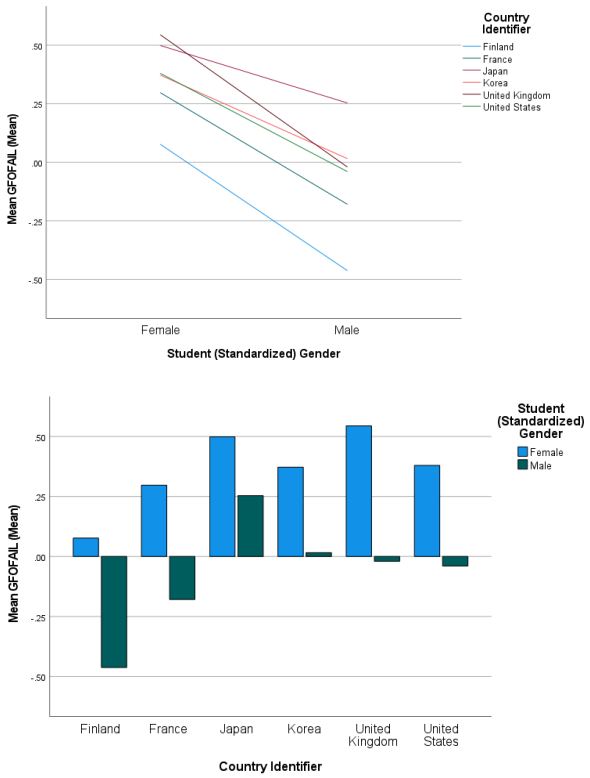


図 6-3：失敗への恐れ

1. 6 か国とも、女子が男子より失敗を恐れる。
2. 日本男子の恐れの高さが、6 か国中際立っている。

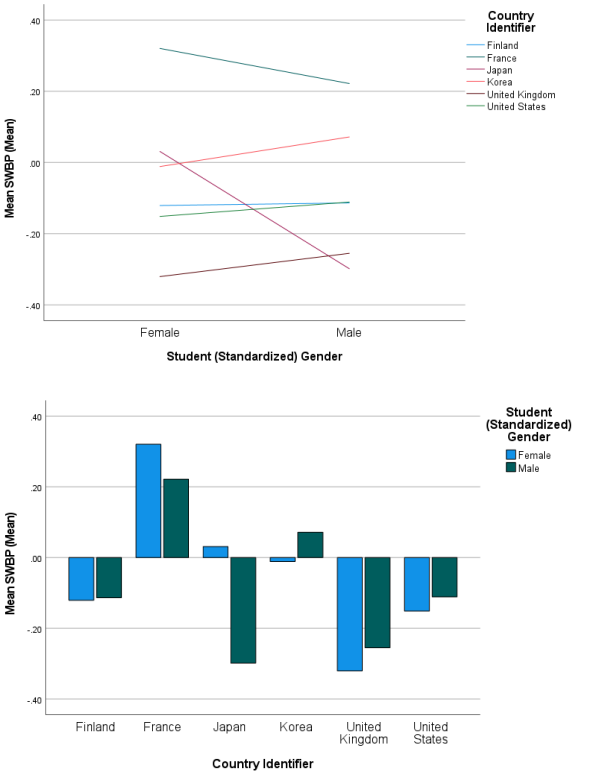


図 6-4：主観的ウェルビーイング（ポジティブな態度）

1. 日本男子の低さが目立つ
2. フランス、日本、英国、米国では、女子が男子より主観的ウェルビーイングが高い（よりポジティブである）。

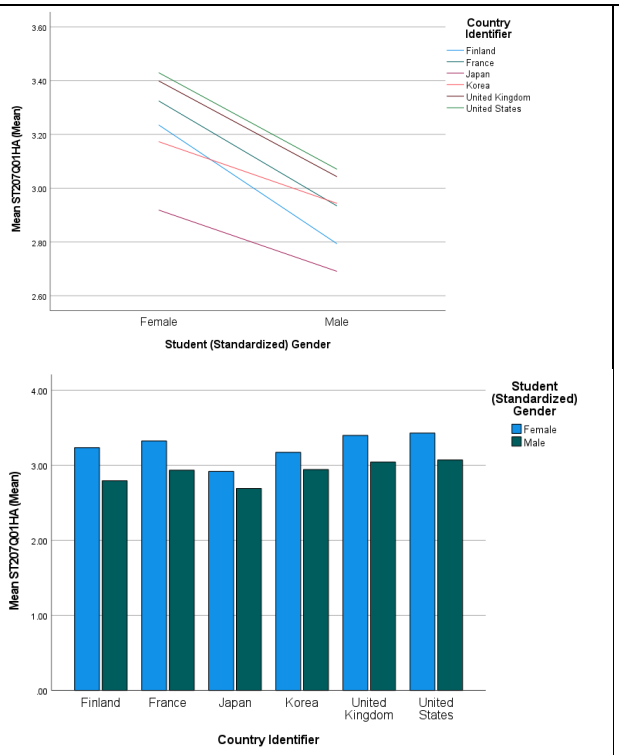


図 6-5：いじめられている生徒を誰も守ってあげないことに腹が立つ

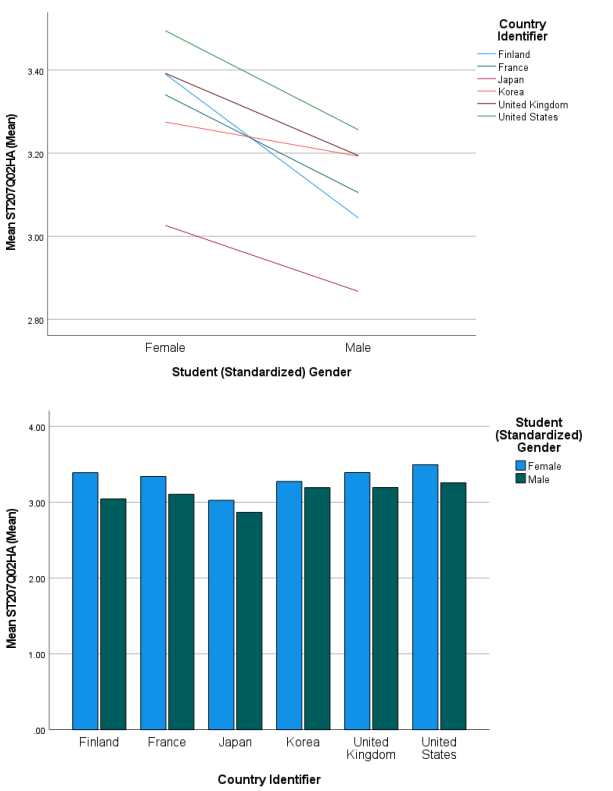


図 6-6：自分を守れない生徒に手助けすることはいいことだ

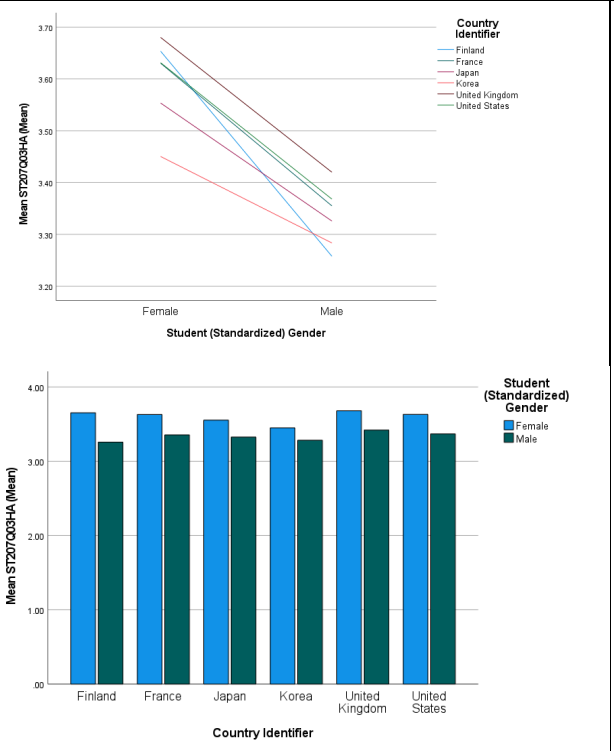


図 6-7：いじめに加わることは悪いことだ

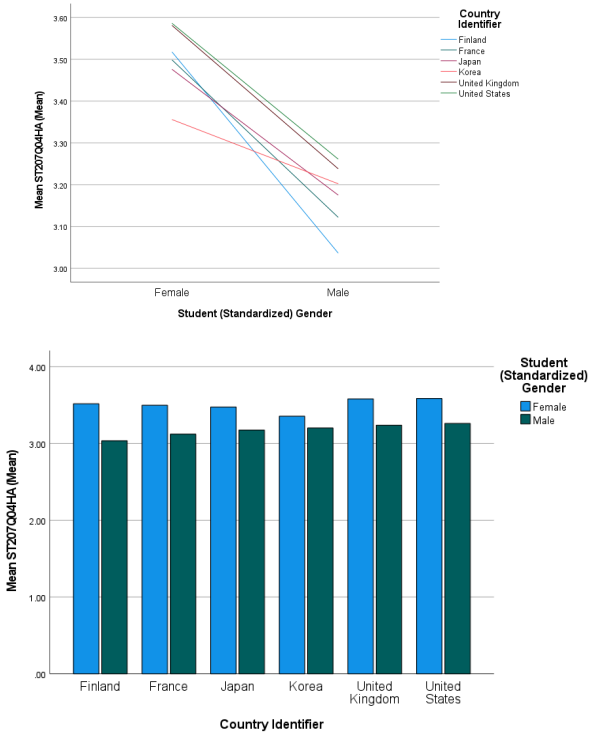


図 6-8：他の生徒がいじめられているのを見るのは不愉快だ

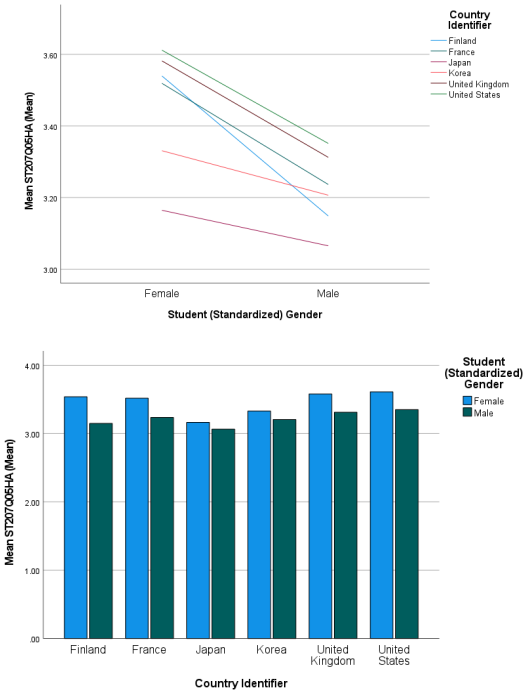


図 6-9：いじめられている他の生徒に誰かが味方するのは、いいことだ

引用文献

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences, 2nd Edition*, Routledge.

Hattie, J. (2023), *Visible Learning: The Sequel*, Routledge.

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.

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.

国立教育政策研究所編(2021)，TIMSS2019 算数・数学教育／理科教育の国際比較―国際数学・理科教育動向調査の 2019 年調査報告書，明石書店.

国立教育政策研究所編(2019)，生きるための知識と技能―OECD 生徒の学習到達度調査（PIISA）2018 年調査国際結果報告書―，明石書店.

熊谷龍一(2012)，統合的 DIF 検出方法の提案－“EasyDIF”の開発－．心理学研究，83.

熊谷龍一(2009)，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.

文部科学省(2022a)， 令和 3 年度「全国学力・学習状況調査」経年変化分析調査テクニカルレポート.（https://www.nier.go.jp/21chousakekkahoukoku/kannren_chousa/pdf/21keinen_tech_01.pdf）

文部科学省(2022b),令和 3 年度『全国学力・学習状況調査』経年変化分析調査テクニカルレポート別冊 （標本抽出方法）
(https://www.nier.go.jp/21chousakekkahoukoku/kannren_chousa/pdf/21keinen_tech_02.pdf)

大久保街亜・岡田謙介(2022)，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（n.d.），講義メモ　古典的テスト理論.
（https://researchmap.jp/sbym_tds/works/37056546/attachment_file.pdf）

田端健人(2024)， 全国学力・学習状況調査「教科に関する調査」の品質検証―平成 28、30 年度、令和 3、4 年度の比較―．宮城教育大学紀要,58.（近刊）

田端健人(2023),「教育の現象学」のデータサイエンス的転回―全国学力・学習状況調査結果の分析から―．学ぶと教えるの現象学研究，20.

田端健人編著(2022)， 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.

Taber, K.S. (2018), The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, 48, Springer.

徳永悠彦（2018），テストは何を測るのか―項目反応理論の考え方―，ナカニシヤ出版.