International Baccalaureate®
Baccalauréat International
Bachillerato Internacional

評価の原則と実践 — デジタル時代における質の高い評価

International Baccalaureate®
Baccalauréat International
Bachillerato Internacional

評価の原則と実践 — デジタル時代における質の高い評価

評価の原則と実践 ― デジタル時代における質の高い評価

2026年5月発行の英語原本『Assessment principles and practices: Quality assessment in a digital age』の日本語版
2026年5月発行

本資料の翻訳・刊行にあたり、文部科学省より多大なご支援をいただいたことに感謝いたします。

発行者　非営利教育財団　国際バカロレア機構（International Baccalaureate Organization）Rue du Pré-de-la-Bichette 1, 1202 Genève, Switzerland
ウェブサイト：ibo.org

国際バカロレア機構（以下、「IB」という。）は、より良い、より平和な世界の実現を目指して、チャレンジに満ちた4つの質の高い教育プログラムを世界中の学校に提供しています。本資料は、そうしたプログラムを支援することを目的に作成されました。

IBは、資料の中で利用する多様な情報源について、情報の正確さと信憑性を確認します。ウィキペディアのようなコミュニティーベースの知識源を使用する際には、特に留意します。IBは知的財産の原則を尊重し、利用する著作物すべてについて刊行前に著作権者を特定し、許諾を得るよう常に努力します。IBは、本資料で利用した著作物に対して許諾をいただいたことに感謝するとともに、誤記および遺漏がありました場合には、可能な限り早急に訂正いたします。

IB の使命

国際バカロレア（IB）は、多様な文化の理解と尊重の精神を通じて、より良い、より平和な世界を築くために貢献する、探究心、知識、思いやりに富んだ若者の育成を目的としています。

この目的のために、IBは、学校や政府、国際機関と協力しながら、チャレンジに満ちた国際教育プログラムと厳格な評価の仕組みの開発に取り組んでいます。

IBのプログラムは、世界各地で学ぶ児童生徒に、人がもつ違いを違いとして理解し、自分と異なる考えの人々にもそれぞれの正しさがあり得ると認めることのできる人として、積極的に、そして共感する心をもって生涯にわたって学び続けるよう働きかけています。

IBの学習者像

すべてのIBプログラムは、国際的な視野をもつ人間の育成を目指しています。人類に共通する人間らしさと地球を共に守る責任を認識し、より良い、より平和な世界を築くことに貢献する人間を育てます。

IBの学習者として、私たちは次の目標に向かって努力します。

探究する人
私たちは、好奇心を育み、探究し研究するためのスキルを身につけます。ひとりで学んだり、他の人々と共に学んだりします。熱意をもって学び、学ぶ喜びを生涯を通じてもち続けます。

知識のある人
私たちは、概念的な理解を深めて活用し、幅広い分野の知識を探究します。地域社会やグローバル社会における重要な課題や考えに取り組みます。

考える人
私たちは、複雑な問題を分析し、責任ある行動をとるために、批判的かつ創造的に考えるスキルを活用します。率先して理性的で倫理的な判断を下します。

コミュニケーションができる人
私たちは、複数の言語やさまざまな方法を用いて、自信をもって創造的に自分自身を表現します。他の人々や他の集団のものの見方に注意深く耳を傾け、効果的に協力し合います。

信念をもつ人
私たちは、誠実かつ正直に、公正な考えと強い正義感をもって行動します。そして、あらゆる人々がもつ尊厳と権利を尊重して行動します。私たちは、自分自身の行動とそれに伴う結果に責任をもちます。

心を開く人
私たちは、自己の文化と個人的な経験の真価を正しく受け止めると同時に、他の人々の価値観や伝統の真価もまた正しく受け止めます。多様な視点を求め、それらを評価し、その経験を糧に成長しようと努力します。

思いやりのある人
私たちは、思いやりと共感、そして尊重の精神を示します。人の役に立ち、他の人々の生活や私たちを取り巻く世界を良くするために行動します。

挑戦する人
私たちは、不確実性に対し熟慮と強い意思をもって向き合います。ひとりで、または協力して新しい考えや方法を探究します。挑戦と変化に、機知に富んだ方法で忍耐強く取り組みます。

バランスのとれた人
私たちは、自分自身や他の人々の幸福にとって、私たちの生を構成する知性、身体、心のバランスをとることが大切だと理解しています。また、私たちが他の人々や、私たちが住むこの世界と相互に依存していることを認識しています。

振り返りができる人
私たちは、世界について、そして自分の考えや経験について、深く考察します。自分自身の学びと成長を促すため、自分の長所と短所を理解するよう努めます。

この「IBの学習者像」は、IBワールドスクール が価値を置く人間性を 10 の人物像として表しています。こうした人物像は、個人や集団が地域社会や国、そしてグローバルなコミュニティーの責任ある一員となることに資すると私たちは信じています。

目次

はじめに

国際バカロレア（IB）のプログラムは、150 を超える国や地域（統治領）で提供され、多種多様な教育的文脈と伝統を包括しています。IB の評価理念や評価方法は、なじみがあるものとして受け入れられる場合もあれば、複雑でなじみがないシステムだと感じられることもあります。

IB は、単に測定が簡単なものを評価するのではなく、本当の意味で重要なものを評価するという信念を貫くことで、その評価原則を通して困難な課題に立ち向かっています。IB の初代事務総長であるアレク・ピーターソンは、教育的エコシステムにおけるこの課題を、バランスの問題であるとしています。

> 必要とされているのは、児童生徒の人生の次の段階を見据えながら、その才能と個性を真に評価するという意味で、できる限り妥当な評価を行うプロセスです。それは同時に、生徒、保護者、教師、進学先・就職先が、評価が正当であると確信できるほどの高い信頼性をもつものでなければなりません。そのプロセスは同時に、その逆流効果によって質の高い指導を妨げたり、適切なタイミングを逃したりしてはならず、ただでさえ不足している教育的リソースを過度に使用するものであってもなりません。
>
> (Peterson, 1971, pp. 27-55)

本資料の目的は、IB が使用している原則を明らかにすることにより、すべてのプログラムにおいて現実に即し、関連性が高く、学習と指導のプロセスを振り返ることができるような方法で、有意義で公平、かつ児童生徒に利益をもたらすような評価が行われていることを示すことです。

IB における評価は大きく分けて、図 1 に示す 2 つのカテゴリーに分類されます。

図 1

総括的評価と形成的評価

目的	**総括的評価:** 習熟度のエビデンスを示す、または生徒の能力を測る。 **形成的評価:** （クラス全体または個人として）改善の余地のある部分や重点を置くべき部分を特定する。
得られるもの	**総括的評価:** 成果物 **形成的評価:** プロセス
評価を実施する時期	**総括的評価:** 学習の集大成として **形成的評価:** 学習期間全体を通して
エビデンスの形態	**総括的評価:** ポートフォリオ、試験、コースワーク（通常は個人で取り組む） **形成的評価:** 観察、退出チケット、2人1組での共有、自己評価（通常は複数で取り組む）
評価ツール	**総括的評価:** 規準やマークバンド **形成的評価:** 発展および学習の進展の足場となるルーブリック

総括的評価と形成的評価のそれぞれの役割を理解することは、IB コミュニティーにとって非常に重要です。

評価の結果は児童生徒の進学や就職に影響し、その人生を左右する力をもちます。評価とは、良い結果と悪い結果のどちらにもつながるツールです。教師、保護者、児童生徒にとって、評価の強み、弱み、そして判断基準を理解することは大きな重要性をもちます。本資料では、以下に示すよくある質問や懸念を取り上げます。

- 「どうして IB はこのような評価を行うのか」 ― IB の評価の実践を裏づける原則を理解する。
- 「この成績はフェアじゃない。もっと良い成績が与えられるべきだった」― 評価判断の公平性と意味を説明する。
- 「これまでそんな風に考えたことがなかった」 ― 評価のプロセスについて解説する。
- 「指導を改善するためだけに評価を使用したい」― 形成的な目的と総括的な目的の両方について、質の高い評価の重要性を強調する。

テクノロジーは日進月歩で進化しています。そのため、児童生徒の取り組み、指導状況、学習方法を本当の意味で反映した評価を実践するためには、IB の評価にテクノロジーを取り入れることは不可欠です。本資料のタイトルにある「デジタル時代における質の高い評価」という言葉が示すように、IB の評価が本物で実用的なものであり続けるためには、児童生徒にとってなじみがあり、アプローチしやすい方法で評価を実施する必要があります。さらに重要な点として、IB がこの転換を行う際は、評価の基盤となっている原則を引き続き遵守する必要があります。これは、評価の妥当性をさらに高めるために、学習と評価をサポートする目的でテクノロジーが使われることを意味します。「デジタル時代における質の高い評価」という表現は、このプロセスおよび時代の変化に対する IB の熱意と対応姿勢を示すとともに、IB の評価の基盤となっている原則を引き続き固持していくという決意の表れでもあります。

本資料は、初等教育プログラム（PYP：Primary Years Programme）、中等教育プログラム（MYP：Middle Years Programme）、ディプロマプログラム（DP：Diploma Programme）、キャリア関連プログラム（CP：Career-related Programme）という IB の全プログラムに適用される IB の評価原則を詳細に説明するものです。この原則には、あらゆる種類の IB の評価が含まれ、それぞれの評価の設計の根拠が明らかにされます。一部のセクションでは、外部採点のために IB に提出される評価課題を含む総括的評価に重点が置かれていますが、本資料に含まれる情報は、IB のプログラム全体の教育実践を強化する目的で提供されています。総括的な外部評価が実施されない文脈においても、何をもって効果的で公正な評価とするかを理解することは、教室での学習体験を改善し、今後の試験に向けて児童生徒の準備を整えるうえで有益です。さらに本資料は、教師、学校管理者、試験官、大学入試課の担当者を含む IB コミュニティー全体を対象に、評価原則を詳しく解説することをねらいとし、IB とその評価について一定の理解はあるものの、その裏にある理論については深く把握していない人々を対象に作成されています。児童生徒に世界最高レベルの教

育体験を提供するうえで、IB の評価原則はその教育理念をどのように支えているのでしょうか。本資料が、その点を理解する一助となれば幸いです。

用語の解説

IB の評価では、教師や児童生徒にはなじみのない用語が多く使われています。すべての IB 資料を通して、正確さを維持しながら誤解を回避するための努力が講じられています。

重要な用語については、初出時に直接説明を加えるか関連資料をリンクしています。

本資料には、評価関連の用語を一覧表記した用語集も含まれています。一般的な評価用語とは見なされない科目に特化した用語は、それぞれの科目の『指導の手引き』で解説されています。

その他のリソース

本資料は、IB の教育アプローチを説明する一連の IB 資料の一部です。重要な IB 資料が互いにどのように関連するかについては、わかりにくい部分もあるかもしれません。表 1 と表 2 は、重要資料の概略を示し、それぞれの目的および役割を説明するものです。

表 1

全プログラム共通の重要な IB 資料

重要な IB 資料 — 全プログラム共通	
IB 資料『評価の原則と実践 — デジタル時代における質の高い評価)』	IB の評価に対する包括的なアプローチと、それがどのように実践されるべきかを説明する資料です。IB 教育の総括的評価（正式な試験）に焦点をあてています。
IB 資料『国際バカロレア（IB）の教育とは』	IB 教育の中核を成す考え方をわかりやすく伝えることを目的とする資料です。IB の全プログラムに共通する教育理念について解説します。IB の教育理念を示すことにより、プログラムの導入に向けた認定プロセスの段階にある学校や、継続的にプログラムを実施している学校に対して、サポートを提供します。
IB 資料『プログラムの基準と実践要綱』	学校と IB の両方が、IB の 4 つのプログラムの実施の成功度を測るための一連の規準を提供する資料です。プログラムの基準（IB プログラムを実施する学校が満たすべき一般的な条件)、実践要綱（基準の詳しい定義。すべてのプログラムに共通)、要件（各プログラムに特化したもの）が含まれています。

表 2

プログラム別の重要な IB 資料

重要な IB 資料 — プログラム別	
IB 資料『PYP：原則から実践へ』 IB 資料『MYP：原則から実践へ』 IB 資料『DP：原則から実践へ』 IB 資料（英語版）『CP: From principles into practice（CP：原則から実践へ)』	特定の IB プログラムの文脈における学習と指導に特化した資料です。プログラムの要件についても説明しています。

重要な IB 資料 — プログラム別	
科目の『指導の手引き』(MYP と DP のみ)	例えば「地理」や「美術」など、1 つの科目、コース、学習分野について詳細な情報を提供する資料。特定の科目のねらい、目標、シラバス、内部評価(IA：internal assessment）の規準を明記しています。これに加え、科目別の学習と指導への指針も提供されます。
「教科のコンティニュアム」(PYP のみ)	学校が「学習範囲と順序」を設計する際に使用できる、参考例としてつくられた資料。地域ごとの文脈や要件に合うように適宜変更を加えることができます。コンティニュアムはまた、各教科とそのストランドに関する概念的理解や概念の例も提供しています。 この資料に付属する教科の概要は、各教科の文脈における PYP の主要要素を明確化し、教科の枠をこえたプログラムにおける教科の役割を明らかにしています。
IB 資料（英語版）『Middle Years Programme assessment procedures (MYP における評価の手順)』 IB 資料『ディプロマプログラム (DP) における評価の手順』 IB 資料（英語版）『Career-related Programme Assessment procedures（キャリア関連プログラム（CP）における評価の手順)』	各プログラムの評価を実施するにあたって守らなければならない規則、規定、特定のプロセスを明記した資料。これらの資料には、IB ワールドスクールが遵守すべき規則の裏づけとなる一般規則が含まれています。
IB 資料『試験の準備に関する方針』(MYP、DP、CP のみ)	学校内における IB 機密資料の安全な保管や試験の日程変更など、試験の準備に関する IB の方針を説明した資料。
IB 資料『試験実施要項』(DP のみ。CP は DP 版を使用）および IB 資料（英語版）『The conduct of IB Middle Years Programme on-screen examinations（中等教育プログラム：コンピューターを用いた試験の実施要項)』(MYP)	コーディネーターと試験監督に対して、各プログラムの試験の実施に関する規定を説明するための資料。各試験会場に必ず 1 冊置いておく必要があります。

これらの概要資料に加えて、教師やコーディネーターなど特定の関係者に固有の手引きを提供する、その他の関連資料も提供されています。このような資料は、IB のウェブサイトおよびプログラム・リソース・センターで入手できます。

本資料の使用方法

本資料で扱われるトピックは多岐にわたり、読み手の関心およびニーズに合わせて、さまざまな使い方ができるように設計されています。このセクションの**トピックの問い**は、特定の読み手にとって重要性の高い側面について学ぶための出発点として使うことができます。

トピックの問い

本資料は、評価に対する IB のアプローチを包括的に説明することを意図したものですが、読み手が評価プロセスの一部について具体的な情報を必要とする場合もあるでしょう。その点を踏まえ、それぞれのトピックの問いの下に、読み手からの具体的な問いに焦点をあてたセクションの一覧を示しています。このリストはあくまでも参考情報であり、必ずしも記載された順番に従って読み進めなければならないということではありません。

関心のある用語を調べるには、巻末の用語解説を活用することもできます。本資料の PDF 版の「検索」オプションを使って、用語解説の中の特定の用語についてさらに詳しく知ることができます。

「IB は評価に対してどのようなアプローチをとっているのでしょうか？」

対象者：IB の評価の根底にある理論、IB における質の高い評価の定義、プロセス設定における意思決定の根拠、相反する評価のニーズの調整方法に関心をもつ大学入試課の担当者、教師、およびその他の関係者

考えられる問いの例：

- IB の評価の目的は何か。何が特別なのか。
- AI は評価にどのような変化をもたらしているか。

上記の問いに特に関連の深いセクションは以下のとおりです。

導入と概要

- 効果的な評価
- カリキュラムの目標を支える評価
- 高い予測可能性
- 幅広い評価課題を活用した評価
- 授業内評価と内部評価の役割
- 協働作業と個別の評点
- 児童生徒のさまざまな能力と高次の思考スキルを考慮した評価
- 高次の思考スキル

- 「IB の学習者像」とスキルの発展
- 国際的な視野と多様な文化への理解
- 優れたデジタル評価とは

セクション A：評価の原則

- 教育における評価
- 評価を定義する
- 評価のアプローチ
- 逆流効果と学習
- 妥当性
- 妥当性を定義する
- 妥当性の議論を構築する
- 妥当性を維持する
- 妥当性の鎖の構成要素
- 妥当性の各側面のバランスをとる
- 信頼性
- 一貫性のある成果と「正しい」成果
- 構成の関連性と現実に即した評価
- 管理のしやすさ
- 公平性とバイアス
- 同等性
- 妥当性に対する IB のアプローチ
- デジタル評価が妥当性にもたらすメリット
- 基準
- 基準の 3 つの側面
- 集団基準準拠と目標基準準拠
- 集団基準準拠
- 目標基準準拠
- 到達度準拠
- 基準の維持
- 総括的評価における生徒の成果の説明
- 成績の影響
- 専門的な判断の重要性
- IB 評価の採点
- 採点の定義
- 採点方法
- 形成的評価の採点

セクション B：IB の評価の実践

- 生徒の到達度を報告する
- IB の成績の意味

- 評点と成績の違い
- 効果的な試験セッション
- 成績と到達度
- 学問的誠実性の定義
- 全員にとっての公正さを優先する

印刷可能な資料

- 「IB の評価の原則」（PDF）

「私は PYP の実践者です。PYP では評価を行わないのですが、この資料は私にとってどのようなメリットがありますか？」

対象者：学校内で実施する総括的評価と形成的評価の機会について、理解したいと考える PYP の実践者

PYP では IB に提出する評価課題はありませんが、学校が独自に実施する総括的評価や形成的評価の機会は、PYP にも存在します。本資料は、それらの領域での優れた実践を紹介しています。また、学問的誠実性や倫理的な考え方を養うことに関する部分も重要です。このような考え方について、年齢に適した方法で早い段階から児童に働きかけていくことが大切です。

上記の問いに特に関連の深いセクションは以下のとおりです。

セクション A：評価の原則

- 評価のアプローチ
- 形成的評価の採点
- 倫理的な考え方の育成
- 評価における人工知能（AI）の倫理的な使用

セクション B：IB の評価の実践

- 学問的誠実性の定義
- 全員にとっての公正さを優先する

セクション C：プログラム固有のプロセス

- 全プログラムに共通の要素
- 初等教育プログラム（PYP）

「私は MYP の実践者です。MYP ではすでにデジタル形式の評価を行っていますが、この資料は私にとってどのようなメリットがありますか？」

本資料には、DP と CP におけるデジタル評価への移行に関する要素が含まれていますが、これらのセクションは、MYP の e アセスメントがもつ実際の価値を実証するものでもあります。本資料は MYP にもあてはまります。トピックの問い、「私は MYP、DP、CP の実践者です。IB に提出する評価課題がどのように採点され、成績が付与されるのかを知りたいと思っています。学校にはどのようなフィードバックが提供されますか？」を参照してください。

「デジタル評価は DP と CP の評価にどのような変化をもたらしますか？」

対象者：IB の評価をよく理解しており、IB においてデジタル評価への移行がどのように行われるかを知りたいと考える教師

考えられる問いの例：

- デジタル評価とは何か。
- 成果物はコンピューターデバイスによって採点されるのか。
- なぜデジタル形式の方がいいのか。
- 同じ基準が維持されるのか。

上記の問いに特に関連の深いセクションは以下のとおりです。

導入と概要

- 効果的な評価
- IB の評価におけるテクノロジーの役割
- 優れたデジタル評価とは
- デジタル評価に関連したリスク

セクション B：IB の評価の実践

- 全員にとっての公正さを優先する
- 評価のライフサイクル
- 評価サイクルに対するデジタル評価の影響

「私は MYP、DP、CP の実践者です。IB に提出する評価課題がどのように採点され、成績が付与されるのかを知りたいと思っています。学校にはどのようなフィードバックが提供されますか？」

対象者：採点のために IB に提出した生徒の成果物がたどる過程と、成績が生成される方法を知りたいと考える教師。また、IB は教師の評価もプロセスの一環としているものの、教師の判断を受け入れない場合もあります。その理由を知りたい教師も対象に含まれます。

考えられる問いの例：

- 今年の成績区分がいつもと違うのはなぜか。
- 別の試験官に採点してもらえないのはなぜか。
- 答案（スクリプト）を採点するのは誰か。
- 試験官はどのようにチェックされるのか。
- 不服申し立てはどのようにすればいいか。
- 自分が付与するであろう評点と違う評点が与えられている。なぜこれが公正といえるのか。
- マークスキーム（採点基準）によれば生徒の解答は評点に値する。なぜこれが公正といえるのか。
- 「成績なしの教室」が生徒にとって最善のアプローチなのではないか。
- なぜ自分の評点が変更されているのか。

- モデレーション係数が年によって変わるのはなぜか。

上記の問いに特に関連の深いセクションは以下のとおりです。

導入と概要

- 授業内評価と内部評価の役割

セクション A：評価の原則

- 教育における評価
- 評価を定義する
- 評価のアプローチ
- 逆流効果と学習
- 妥当性の鎖の構成要素
- 成績の影響
- IB 評価の採点

セクション B：IB の評価の実践

- 生徒の到達度を報告する
- IB の成績の意味
- 評点と成績の違い
- 効果的な試験セッション
- 成績と到達度
- 予測スコア
- 評価のプロセス：役割と責任
- 主任試験官と試験官長
- その他の試験官の役割
- IB スタッフの責任
- 試験官の序列
- 全員にとっての公正さを優先する
- 評点と成績
- 採点に対するアプローチ
- 分析的マークスキーム
- 包括的な規準：マークバンド
- 標準化
- 品質モデル
- 練習
- 認定
- ライブマーキングとシード
- 効果的な標準化の指標
- 許容差
- 判断に迷う答案および通常とは異なる答案
- 学校のつながり
- 設問項目グループ

- 集約
- モデレーション（評価の適正化）
- モデレーションと教師に対する期待事項
- 生徒の成果物の選定
- 例外となるケース
- モデレーション係数を導き出せない場合
- ダイナミックサンプリング
- 成績の付与と集約
- 判断に基づく成績区分と補間的な成績区分
- 成績付与のプロセスで使われるエビデンス
- 今年の受験者群を考慮する
- 評価についてのフィードバック
- 答案のエビデンスを確認する
- 結果の統計を確認する
- エビデンスのバランスをとる
- 固定された成績区分
- 推奨された成績区分の確認と承認
- プログラム認定証の授与
- 教師オブサーバー
- 成績付与プロセスの原則
- 品質確認
- 学校へのフィードバック
- 内部評価のフィードバック

セクションC：プログラム固有のプロセス

- 全プログラムに共通の要素

「IBの試験はどのように設計・作成されていますか？」

対象者：試験の作成方法を知りたい教師、ならびに試験の質について懸念を抱く教師

考えられる問いの例：

- 試験問題を作成するのは誰か。
- どのように確認されているか。
- 試験問題は異なる言語間で同一なのか。
- 正しい内容をテストしているのか。

上記の問いに特に関連の深いセクションは以下のとおりです。

導入と概要

- 効果的な評価
- カリキュラムの目標を支える評価
- 高い予測可能性
- 幅広い評価課題を活用した評価

- 授業内評価と内部評価の役割

セクション A：評価の原則

- 教育における評価
- 評価を定義する
- 評価のアプローチ
- 逆流効果と学習
- 妥当性
- 妥当性を定義する

セクション B：IB の評価の実践

- 試験の作成における役割
- 全員にとっての公正さを優先する
- 評価の作成
- 試験制作の概要
- 内容の承認までのプロセス
- 作成と内容の承認
- 組体裁と校正
- 使いやすさの承認
- 品質管理
- 翻訳
- 調整済み試験問題
- 評点と成績

評価の原則と実践

評価の原則とは、認定のための試験や評価を作成し、提供し、採点し、成績をつけるうえで重要だと見なされる事柄です。これらは、IB の教育について私たちが重要だと考えるものに由来しています。最も重要な原則は、評価とは教育を歪めるものではなく、教育を支えるものであるべきだということです。

評価の実践とは、IB がその原則を有意義かつ実用的な方法で実行することです。信念を貫くという IB の理念を維持しながら、世界中に存在する相反するニーズや実践に関する制約が考慮されています。

「IB の評価の原則」の要点は以下のとおりです。

評価とは、総括的評価と形成的評価の別なく、以下を満たすものでなければなりません。

- 意図された目的に対して妥当であること。これは、「構成の関連性」（正しい内容をテストしているか）、信頼性、公平性（バイアスがないか）、代替手段との比較、生徒、学校、IB にとっての管理のしやすさという相反するニーズを、バランスよく満たす必要があることを意味します。
- プラスの逆流効果をもたらすこと。つまり、質の高い学習と指導を促す設計になっていること。
- 個人的な到達度を実証するという意味で、できる限り大勢の生徒にとって適切な内容になっていること。
- あらゆる生徒が評価を受けられるよう、必要な受験上の配慮をすべて実施できるように設計されていること。
- 個々のプログラムを個別に考えるのではなく、IB プログラム全体の文脈にあてはめられること。
- 学習の同時並行性と、生徒の全体的な学習活動を支えられること。
- IB の幅広い理念を支え、探究する人、知識のある人、考える人、コミュニケーションができる人、心を開く人という学習者像を発展させられること。

評価の実践は通常、概要として説明され、個別の科目、教科、プログラムの文脈の中で適切に実施されます。評価の実践は、日々のプロセスの詳細や、プログラムや科目ごとの実施状況の違いに言及することなく、実施しなければならないことを大まかに説明したものです。

教師向けの印刷可能な資料として「IB の評価の原則」（PDF）が用意されています。

効果的な評価

- 「効果的な評価」を表す単一の定義というものは存在せず、その意味は、何を優先順位とするかによって、また、所定の評価の目的によっても変わります。
- IB の評価の根底にある原則は、評価可能なものを評価するのではなく、重要なことを評価する、ということです。
- IB の評価では、学習と指導にプラスの逆流効果をもたらすことを目指しています。
- 概して、効果的な評価には、生徒の学習の各側面をバランスよく捉えたさまざまな評価課題が含まれていなければなりません。また、教室での学習活動の掘り下げや試験の機会を含む必要もあります。
- 学習のスタイルや好みがさまざまに異なるすべての生徒が評価を受けられるようにするには、評価の原則を念頭に置きながら、ユニバーサルデザインを採用した評価を設計・作成する必要があります。関与、参画、行動、表現について、さまざまな手段を検討する必要があります。

IB が考える**効果的な評価**とは、**有意義な学習成果を重視**する評価のことです。評価をカリキュラムの目標と密接に整合させ、生徒が幅広いスキルと能力を発展させられるようにサポートします。

IB の評価の根底にある原則は、評価可能なものを評価するのではなく、**重要なことを評価する**、ということです。この原則は、信頼性や生徒の作業負担といった他の検討事項とのバランスを考慮しながら実現する必要があります。

単一のアプローチのみで、すべての重要事項を適切に実現することは容易ではありません。特に、効果的な評価の設計は、総括的評価と形成的評価との間で異なります。この原則を踏まえて、IB は効果的な評価を以下のように定めています。

- カリキュラムの目標をサポートする
- さまざまな種類の評価課題を活用する
- 児童生徒のさまざまな能力と高次の思考スキルを考慮に入れる

上記のそれぞれについて、以下のセクションで詳しく見ていきます。

カリキュラムの目標を支える評価

- 評価は学習と指導にプラスの逆流効果をもたらし、効果的な学習と指導を促す必要がある。
- 評価は高い予測可能性をもつ必要がある。

評価は、学習および指導と切り離して考えるべきではありません。IB の評価結果は総括的評価に基づいており、学習と指導に対して直接的なフィードバックを提供することを意

図したものではありません。ただし、評価を構成する要素が指導内容に影響を与えるということが広く理解されており、これは「逆流効果」として知られています。

IB にとって、評価の設計とは、すべての児童生徒に対して最も望ましい教育成果をもたらすものであるべきです。評価が児童生徒の学習にもたらす影響は、IB の評価の設計において依然として重要な検討事項となっており、構成の関連性と併せて、妥当性のさまざまな側面を適切なバランスで実現する助けとなります。

重要な評価が指導と学習に及ぼす影響は、効果的な指導と、生徒が自身の学習に建設的に関与することを促すような評価を設計することにより、有効に利用することができます(Murphy, 1999 などを参照)。

IB プログラム全体を通して、評価は学習活動と深く結びつき、その成果を振り返る役割を果たします。評価はすべて、建設的な整合のアプローチ（Biggs, 1996）を使いながら、意図された学習成果を中心に設計されています。このような学習成果は、児童生徒への指導と評価の方法を形づくるとともに、その指針となり、成功のための準備を確実に整えられるようにしています。詳しくは、IB 資料『IB における評価のアプローチ』を参照してください。

「IB の使命」で表現され、「IB の学習者像」で説明される、児童生徒が目指すべき資質は、**児童生徒の学習に関する構造主義的理論**と非常に高く整合しています。この枠組みにおいて、児童生徒は積極的に学習プロセスに関わり、自らの学習の責任を負い、探究を通して知識、理解、スキルを発展させていきます。

例えば、自分の文化以外の文化的観点に、開かれた心で積極的に関わることは、複数の科目の評価要件において直接的に明示されている期待事項です。より情動的な性質をもつ思いやりや共感を正式な評価に含めることは比較的難しいですが、このような資質も全体的な評価システムに含まれていなければなりません。これは、PYP の「エキシビション」、MYP の「コミュニティー活動」、DP の「創造性・活動・奉仕」（CAS：creativity, activity, service）の要件、および CP の「コミュニティー活動」のような、カリキュラムにおいて評価の対象にならないプログラム固有の要件を通して部分的に達成されます。

IB では、評価が実施される方法が学校内での IB のコースの指導方法に大きな影響を与えるという認識の下、特に DP および CP コースの設計に関して、予測可能な妥当性（結果が将来の成功を予測する度合い）に重点を置いています。各科目に適用される評価モデルは、幅広さをもち、さまざまな種類のエビデンスを包括するとともに、生徒の到達度と学習を裏づける多様なエビデンスを提供することで構成の関連性を支えられるように設計されています。

IB は、評価が学習と指導に逆流効果をもたらすことを十分に認識しながら、同時に、IB プログラムのすべての目標と理念を児童生徒の中に浸透させられるような指導方法を採用するよう学校に働きかけています。

IB では定期的に研究調査を実施し、それによって、将来の学習に向けて効果的に準備を進められるようなプログラムを設計することに成功しています。IB が実施したさまざまな調査研究について詳しくは、IB ウェブサイトの「Research（調査研究）」のセクションを参照してください。

高い予測可能性

「予測可能性」とは、将来何が起こるか、また、特定の出来事がいつ起こるかを予測できることです。評価における予測可能性とは、特定の試験において、どのタイミングで、どのような問題が出題されるかを学校が予測できることを指します。評価における高い予測可能性は、試験の設問を、教師が生徒の指導に使うカリキュラムと整合させることにより、公平性を生み出します。そして生徒は、細かい内容までは正確に予測できないものの、大まかにどのような種類の問題が出されるかを理解することができます。

IB は、特定のカリキュラムが改訂されるまでの期間全体を通して、学校が指導に費やした時間と労力が報われるようにしたいと考えています。該当する評価の方法に合わせて、シラバス全体を精査する必要があります。予測可能性を下げるような状況を排除するために、コースの終盤に多大な注意を払う必要があります（その時点でまだ出題されていない内容は、試験に出る可能性が高くなります）。

根底にある原則は、生徒にとっても学校にとっても想定外の問題が、試験に出題されるべきではない、いうことです。したがって、どの評価要素においても、試験で問われる内容は科目の『指導の手引き』に直接示されているものでなければなりません。評価においては、予測可能性がもたらす悪影響を可能な限り低減することを目指しています。そのために、事前に解答を準備した生徒が有利な立場に立つことがないよう、試験を設計しています。

予測可能性の低下を抑えるには、所定のテーマ、選択項目、テクストを試験するための方法を十分に用意することが必要ですが、それには評価の設計が非常に重要となります。

表 3

高い予測可能性と低い予測可能性

予測可能性が高い	予測可能性が低い
テーマ、選択項目、テクストおよび指示用語に基づいて問題のあらゆる組み合わせが検討された場合、当該科目に関して IB が出題する問題が、その結果に含まれることになる。	教師が、複数の試験セッションにわたって繰り返し出題されている問題があることを認識し、生徒の試験対策としてそのような問題に過度に集中的に取り組ませる。これは、複数のスキル（知識、分析、評価）を示すことが求められる論述形式の問題に支障をきたすことになる。
問題の再利用に対して、予測できないアプローチをとる。	評価の設計に関する意思決定が適切ではなかったことにより、出題する問題の範囲が限られたものになる（例えば、出題の対象となるメインテーマが 1 つしかないコースにおいて、規定のテクストを設定する）。つまり、設計によって予測可能性が悪化することになる。

幅広い評価課題を活用した評価

多肢選択問題、短答式問題、論述式問題、小論文、プロジェクト、ポートフォリオ内の 1 つの成果物、研究課題はすべて評価課題の例に含まれます。

評価方法または評価要素は、1 つの課題で構成されることもあれば、テーマ、内容の継続性、利便性などに基づいて組み合わされた複数の課題で構成されることもあります。評価方法や評価要素の例として、総括的な試験、生徒の成果物を集めたポートフォリオ、長期的な協働プロジェクト、研究課題などが挙げられます。評価課題の概念と評価要素の概念には重複している部分があります。場合によっては、1 つの評価要素に対して、複数の選択肢から 1 つの課題を選択して取り組むこともあります。

IB が多岐にわたる評価課題や評価要素を使用するのには、いくつもの理由があります。まず初めに、歴史的かつ実用的な観点から、ピーターソン（Peterson, 2003）は DP の評価の初期の発展について、「多くの IB の生徒が入学を希望する教育機関を擁する国や地域（統治領）で使用されている、さまざまな評価手法を考慮する義務があり、そのような評価手法を考慮する機会もあった」と述べており、この原則は MYP と CP にも適用されます。また、目的に適うかどうかという点に関して、妥当性をめぐる検討事項もあります。妥当性を実現するには、さまざまな評価のアプローチをとる必要があるためです。最後に、多岐にわたる評価手法を使用することは、評価において不公平性が生じる可能性を低減することにつながります（Linn（1992）および Brown（2002）を参照）。幅広い評価要素を使用し、その中で課題を設定することにより、科目全体の評価モデルを俯瞰した場合に、生徒の到達度が当該科目のすべての目標に対して適切に表されるようになります。

授業内評価と内部評価の役割

授業内評価は、筆記試験に適さない分野において生徒を評価する機会を数多く提供します。最も重要な側面として、通常の試験では時間の制約によるプレッシャーが存在しますが、教室を基盤とした評価では、長期的な課題を通して、生徒が時間的なプレッシャーを感じることなく問題を調査し、自らの思考の発展を示す機会を得られるという点が挙げられます。これはつまり、授業内評価としてしか実施できない、多種多様な評価課題が存在するということです。

授業内評価として実施すべき課題の例として、プロジェクト、フィールドワーク、実験室での実習、パフォーマンス、数学的研究などが挙げられます。

内部評価には、国際的な資格という文脈において、他にも複数の**利点**があります。他の評価と同様のスキルに対応しながらも、幅広いトピックを選択できる**柔軟性**をもつため、学校は、**地域、文化、地理的な文脈**の中に学習を位置づけたり、教室と実社会とをより密接に結びつけたりすることが可能になります。学校が所在する場所とは異なる文化的背景をもつ生徒が集まるインターナショナルスクールなどでは、地域の社会や環境とより深く関わるための手段として、内部評価を利用することができます。

また、内部評価では多くの場合、それぞれの生徒が自分でトピックや課題を選択することができます。これにより、自分が特に関心をもつ事柄を掘り下げ、より主体的に自身の

学習に取り組むことができます。内部評価は、その柔軟なアプローチによって生徒の教育に付加価値をもたらし、評価プロセスだけでなく、学習活動全体の有効性を高めます。

ただし、授業内評価には**課題**も存在します。中でも特に重要なのが、**学問的誠実性**の確保です。例えば、授業の外で課題に取り組むなど、標準化された状況（監視の目がある状況）以外で課題を完成させた場合、IB の学問的誠実性に関する方針に反する行為が行われなかったことを確認するのは非常に困難です。これには例えば、第三者に課題の作成を依頼することなどが挙げられます。また、インターネットに誰もがアクセスできるようになったことで、この問題は飛躍的に複雑化しています。このような理由により、IB では、生徒のコースワークが本人によるものであることを確認するとともに、生徒が自分で取り組んでいない成果物が提出された場合に、それを最も正確に特定できるのは、学校の教職員であると考えています。学問的誠実性に関する詳細および関連資料については、IB ウェブサイトの「Academic integrity（学問的誠実性）」のページを参照してください。

授業内評価のもう 1 つの課題は、教師と生徒に大きな**作業負担**を生じさせる可能性があるということです。授業内で課題を実施した場合、指導に使える時間が削られることになります。また生徒にとっても、内部で設定された課題は通常、完成までに多くの時間を必要とします。内部評価に求められるスキルやプロセスの練習をするために時間を割くことは適切ですが、教師も生徒も、内部評価のために設定された特定の課題のリハーサルや練習に必要以上の時間を費やしがちになり、それがさらに指導時間を削ってしまうことになります。DP 科目の『指導の手引き』において、内部評価を支えるうえでの教師の役割について詳細なガイダンスが提供されています。

授業内の課題は、**内部**または**外部**で評価されます。外部評価の場合、生徒が完成させた成果物は IB に送られ、試験官によって評価されます。試験官の判断の質は、IB によってモニタリングされます（本資料の「採点」のセクションを参照）

一方内部評価の場合、学校（通常は生徒の担当教師）が生徒の成果物を採点し、IB がモデレーション（評価の最適化）のプロセスを実施して、グローバル基準が正しく適用されたかどうかを確認します（本資料の「モデレーション」のセクションを参照）。

内部評価に関する見方は、世界各地で異なります。1 つの評価で「断片的に」評価するのではなく、生徒の学習状況を包括的に理解するうえで、最も適した立場にある教師の意見に大きな重点を置いている教育制度もあれば、生徒の成果に基づいて教師の指導能力を評価している教育制度もあります。この場合、教師が内部評価に対して高い評点を与える動機となります。

教師の判断は、生徒の過去の成果物の質に影響を受けることもあり、これによりさまざまなバイアスが発生する可能性があります。教師は、内部評価に取り組む生徒の指導とサポートにおいて、自分たちが果たすべき役割の範囲を明確に理解していないことがあり、担当科目における到達度のグローバル基準について限られた見解しかもっていないことも多くあります。教師はまた、自分の生徒の成果物を評価する際に、学校内に存在する一般的な基準に大きく影響を受けることもあります。評価のために提出された課題に対して過度に高い評点を与える動機がない場合でも、教師という仕事を通して築いた生徒との関係が客観的な判断を難しくすることもあります。このような無意識のバイアスに関する研究

では、それが顕著な影響を及ぼす可能性があることが示唆されています（例えば、Zanga, De Gioannis, 2023 を参照）。

IB は、生徒の成長を見守ってきた人物によって評価される授業内の成果物という形式を通して、有意義な評価課題を設定することで得られるメリットは、内部評価がもたらすリスクよりも大きいと考えています。したがって、内部評価課題は通常、科目の評価モデルに不可欠な要素とされます。

協働作業と個別の評点

複数の生徒が協力して 1 つの活動に取り組む協働的な課題を評価することには難しさが伴います。これは将来、生徒が職場において直面することになる重要な要素ですが、学問的な評価においてはしばしば見落とされています。

状況によっては、グループ作業における各個人の役割を特定することは可能です。例えば、ダンスパフォーマンスなどでは、一人ひとりのスキルを見て個別に採点することができます。このような状況では、グループ作業は大きな問題にはなりません。

一方、最終的な成果物において、誰がどのように貢献したかを判断できない場合もあります。このようなケースではグループ全体に 1 つの評点が与えられることがありますが、この方法は、最終的な成績に対する各生徒の貢献度の違いを考慮していません。例えば、1 人の生徒が成果物の大部分を完成させ、深い知識を発揮したとしても、他の生徒より高い評点を得ることはありません。IB の成績は一人ひとりの生徒に個別に付与され、それが選考プロセスに用いられることになるため、この方法は公平なアプローチではないと考えます。IB では原則として、生徒の到達度を個別に測定できないグループ評価は避けるようにしています。

児童生徒のさまざまな能力と高次の思考スキルを考慮した評価

IB 教育では、単なる「事実」の学習をこえた成果の達成を目指しています。IB が長期にわたって掲げてきたこの目標は、21 世紀型の**スキル**、**職場で役立つ能力**、またはそれに類似する**イニチアチブ**を児童生徒に提供する必要があるという、各国政府の現在の考え方に反映されています。

生徒の能力に関する IB の理念とアプローチは、**「IB の学習者像」**および**国際的な視野**へのつながりに焦点をあてています。この点は、プログラム間のつながりについて述べた「セクション C：プログラム固有のプロセス」で詳しく説明されています。心に留めておくべき最も重要な概念は、IB の評価アプローチでは、生徒は学習内容に関する知識だけでなく、現実社会での応用を反映するさまざまな能力に基づいて評価されるということです。IB の評価は、慎重な設計を通して、将来の課題に対応する能力を備えた**生涯学習者**を育てることをねらいとしています。

したがって、優れた評価とは、各コースが達成しようとしている**あらゆる成果**を考慮に入れ、生徒がそのすべてにおいて能力を発揮できるような評価です。ただし、これらの結果の一部のみを測定することが望ましい場合や、管理上の制約により一部の測定しか実施

できない場合が多く、優れた評価とは、これらの制約のバランスをとるものでもあります。IB が達成しようとしている成果は、**高次の思考スキル**、生徒のさまざまな**能力**、そして**国際的な視野**へのつながりに分類されます。

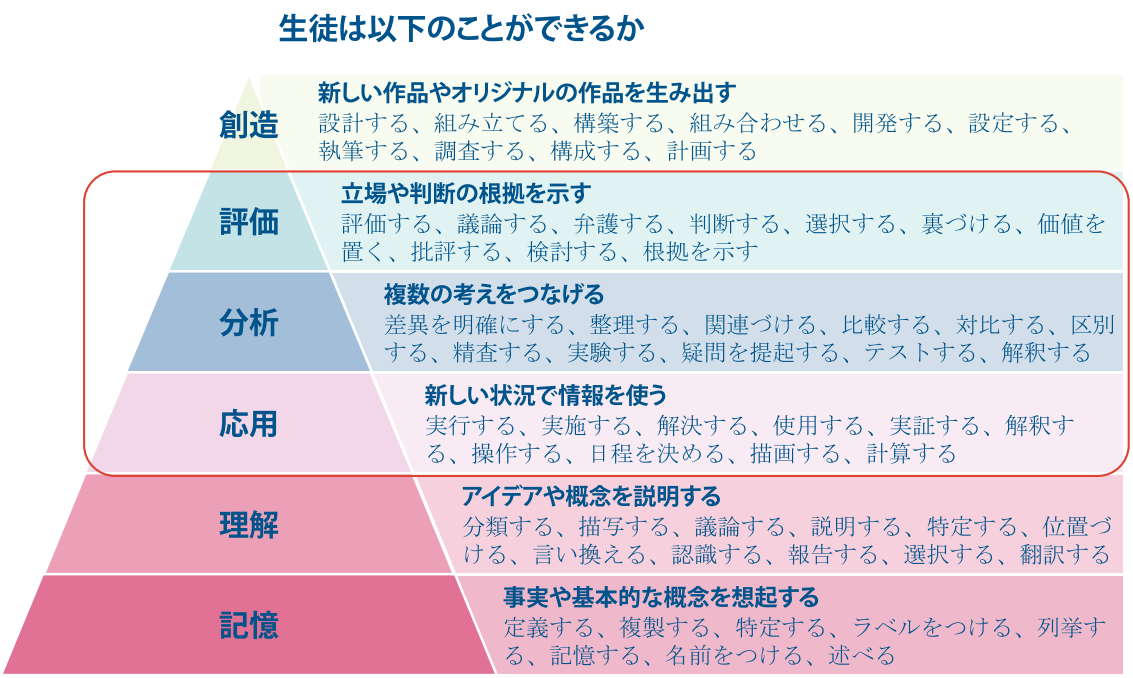
高次の思考スキル

IB の評価は、単なる知識の暗記ではなく、**評価**や**分析**といった高次の思考スキルを測定することを目指しています。

アレク・ピーターソン（Peterson, 2003）が提唱したこの観点は、IB の教育理念を形づくっています。ピーターソンは、「事実、または事実からの解釈を取り入れて反芻することが重要なのではありません。新たな状況や事実が提示されたときに応用できる精神力や考える力を発達させることこそが重要です」と述べています。2011 年にシンガポールで開催された IB 校長カンファレンスで講演したスガタ・ミトラは、ピーターソンの視点をさらに一歩進め、インターネットを通じて包括的な知識がリアルタイムで入手できるということは、知識自体にほとんど価値がないことを意味し、むしろ 21 世紀の市民に必要なのは知識を分析し、解釈し、創造する能力であると主張しています（Mitra, 2011）。

IB の評価では、例えばブルームの分類法（Bloom et al., 1956; Anderson, Krathwohl, 2001）などの確立された分類法に基づいて評価課題におけるパフォーマンスの期待値を設定するなど、高次の思考スキルに大きな重点が置かれています。実際、教育目標に関するブルームの分類法は、評価課題を適切に完了し最高点を獲得するために必要なスキルと認知プロセスの多様性を表現する、有益な枠組みを提供します。ブルームの高次のスキルは、さまざまな種類の評価の使用を求めるものです。生徒の分析スキル、統合スキル、評価スキルは、ある程度詳しい分析、統合および評価を生徒に求めることによってのみ、適切に測定することが可能になります。パフォーマンス評価は、このような領域における生徒の到達度を測定する唯一の現実的な手段です。その成果は厳密に規定できないため、多種多様な正しい解答を許容する、比較的構造化されていないオープンエンド型の評価でなければなりません。

図 2
ブルームの分類法



私たちの評価を有意義なものにするためには、これらのスキルにおける生徒のパフォーマンスを認識し、それに見合う評価を与える能力が非常に重要です。ただしこれは、信頼性などを含む妥当性の各側面に対して課題を提起することになります。知識、概念および定型的な手法を思い出すことしか重視されないテストは、高い重要性をもつ高次の思考スキルに十分に対応していないため、IB 教育の目標において目的に合うものとは言えません。

「IB の学習者像」とスキルの発展

> 今日の教育は、問題解決と意思決定に対する創造的で批判的なアプローチを含んだ考え方を非常に強く重視しています。さらに、コミュニケーションや協働などの取り組み方に加えて、新しいテクノロジーの可能性を認識して利用する能力や、新しいテクノロジーのリスクを実際に回避する能力など、そのような取り組み方に必要なツールも重要視されています。最後に、教育は、能動的で積極的に関与する市民として多面的な世界を生きるための能力にも焦点を合わせています。このような市民は自分が学習したいことや求める学習方法に影響を及ぼし、それが教育者の役割を形づくることになります。
>
> (Schleicher, 2016)

近年、21 世紀に必要とされるスキルは過去の世代と根本的に異なるという主張が強まっています。これらの 21 世紀型スキルを支える探究のアプローチはソクラテスの頃から重視されていたという議論もありますが、人生に備えるための幅広い資質を生徒に与えることの重要性については一般的な合意があります。例えば、ルウェリン (Llewellyn, 2014) の主張などを参照してください。

21 世紀型スキルを分類する方法には、経済協力開発機構 (OECD) の 21 世紀の能力、RAND Education、全米研究評議会 (NRC) の枠組みなどさまざまなものがありますが、IB では、「IB の学習者像」においてこれらの能力を定義・説明しています。

「IB の学習者像」のすべての側面が総括的評価を通した測定に適しているわけではありませんが、そのうち複数の側面が高次の思考スキルの概念に組み込まれています。効果的な評価とは、このような資質の重要性を認識し、形成的評価と総括的評価を通して、該当する能力を発展させるための空間を作り出すものです。その例として、「心理学」や「理科」の領域でアンケートと実験で倫理的なアプローチを推進すること、「振り返りプロジェクト」において倫理的なジレンマの探究を発展させること、適切な相互評価をサポートすること、そして予想外の文脈を生徒に紹介することなどが挙げられます。

生徒の能力を育成する IB の幅広いアプローチについての詳細は、IB のウェブサイトから入手できる資料または各プログラムの IB 資料『原則から実践へ』を参照してください。

国際的な視野と多様な文化への理解

「IB の使命」は、生徒が「多様な文化への理解と尊重の精神を通じて、より良い、より平和な世界を築くことに貢献する思いやりに富んだ若者」「人がもつ違いを違いとして理解し、自分と異なる考えの人々にもそれぞれの正しさがあり得ると認めることのできる人」として成長する手助けをすることです（「IB の使命」、2024）。IB プログラムは、多数の国や地域（統治領）で、多数の国籍をもつ生徒に提供されています。したがって、IB の学習と指導には**国際的な文脈**と**国際的な視野**への焦点の両方が存在し、評価にはその両方が反映されていなければなりません。この複数のものの見方を実現するための重要なステップは、評価の作成者やカリキュラムの開発者など、幅広い文化的背景をもつ学識経験者を擁することです。「IB の使命」を実現するには、違いを曖昧にするのではなく、生徒が生産的かつ積極的に違いを探究できるようにすることが大切です。

一部の教科では、カリキュラムの中でさまざまな文化的観点を認識し、文化的に異なる内容に焦点をあてることを通して、文化的多様性を推進することができます。このアプローチの例は、「生物」「化学」「心理学」「美術」などの DP 科目において見られます。このうち最初の 3 科目は選択項目を含むように構成されているため、学校は、その科目を指導するにあたっての特定の文化的伝統にある程度適合するような学習内容を選択できます。

「芸術」や「言語と文学」などのその他の科目では、カリキュラムにおいて推奨また規定された学習内容を通して、国際的な視野を推進することができます。また、さまざまな内部評価課題を通して、国際的な視野を強調することもできます。

国際的な視野には、他の文化についての単なる知識や理解以上のものが含まれます。**態度**と**行動**も重要な資質です。通常の学校の評価を通して態度を評価することは容易ではありません。

IB プログラムでは、MYP の「コミュニティープロジェクト」や DP の CAS のような評価対象外の「コア」要素を通してこの課題に取り組みます。プログラムの「コア」要素を完了していない生徒には、MYP 修了証、IB ディプロマ、および IB の CP 修了証は授与されないため、これらの評価対象外の要素は IB の評価の全体的な結果に重大な影響を及ぼします。

図 3
信念のある行動

評価においてさまざまな国際的要件に対処するためには、解答する問いを受験者に選ばせることが最善の方法に思えるかもしれません。ただし、この方法では、評価において選択項目間の**同等性**を維持するという点で問題が生じます。この問題は、問いの選択肢が複数用意されている場合や極めてオープンエンド型の評価課題で発生することが多く、生徒の教育的バックグラウンドがさまざまに異なる状況において、「同等の要求度」の意味を定義することをさらに難しくしています。一般には、生徒が自身の経験を解答に組み込むことができるような共通の課題を設定することで、同等性の維持が容易になりますが、その場合、試験官が共通の基準を維持することが難しくなります。

同等性の情報は生徒のパフォーマンスの分析から収集することが可能です。この分析については「成績の付与と集約」のセクションで詳述します。

国際的な文脈で実施される評価は、同一国内での評価で通常発生する問題に加えて、公平性に関する難しさを伴います。ある国の状況では完全に適切な問いが、別の国では不適切な問いになる可能性があります。例えば、スポーツ、旅行、エンターテインメント、歴史上の出来事に言及する問いはもちろん、天気に言及する問いであっても、細心の注意を払って作成しなければなりません。この問題を回避する唯一の方法は、社会文化的文脈を排除した試験問題を作成することであるように思えるかもしれません。ただし、そのような試験問題は非常に限定的なものになるだけでなく、IB の評価についての理念全体に矛盾するとともに、文脈に基づいた課題を通して妥当性を保証するという優れた評価の実践にも反することになります。文脈化された取り組みと評価は、優れた学習に不可欠です。

このジレンマを解消する方法は 2 つあります。1 つ目は、科目のシラバスの内容を具体化することを通して、適切な文脈情報を生徒に提供することです。設問の前提となるケーススタディーを示したり、場合によっては、この情報を試験の内容そのものに含んだりすることもできます。

2 つ目は、生徒が自身の文脈を使って解答できる、よりオープンエンド型評価の設問と課題を用いることです。この方法を使う場合は、内容に共通基盤がないため、科目の内容についての単純な知識ではなく、より深い理解度に採点の焦点を合わせることが重要になります。これは IB の評価理念によく沿った方法です。

図 4

文化的文脈の範囲

上記 2 つの方法を両方採用した場合でも、生徒が自身にとってなじみのない社会文化的な文脈を取り上げた評価課題に直面する可能性があります。これは、**より柔軟な態度**を示し、**グローバルな認識を深め**、なじみのない文化的文脈においても**臆することなく、能力を発揮できる**ような人物像の育成を目指す IB の評価理念に整合するものです。

IB の生徒の多くが、母語または自分が最も得意とする言語以外の言語で試験を受けます。その大部分が英語で試験を受けるケースであり、フランス語またはスペイン語（IB の評価が実施される他の 2 つの言語）で試験を受ける生徒は、その言語の流ちょうな話者であることがほとんどです。したがって、評価の作成と校閲を行う際は、設問の文章が付加言語の話者に不利になるような書き方になっていないことを慎重に検討します。これは、試験の編集作業において重要な優先事項とされています。

IB の評価におけるテクノロジーの役割

- 生徒は、日常での他者とのやりとりでも、学習に取り組む際や論文を執筆する際にも、さまざまなテクノロジーを使用します。そして進学先や就職先でも、引き続き多くのテクノロジーを使用することになります。したがって、試験におけるテクノロジーの使用は、実社会における経験に即したものといえます。
- IB では、有意義で質の高い評価を実施するためにテクノロジーを使用しますが、テクノロジーの使用が評価に対するアプローチの選択に影響することはありません。
- 専門知識をもつ試験官をサポートするためのテクノロジーの使用（e マーキング）の影響と、生徒にとって有意義な評価を作成するためのテクノロジーの使用は、分けて考えることが大切です。
- 評価のために生徒が使用するテクノロジーは、普段の授業で使っている、慣れ親しんだものでなければなりません。

デジタル評価

デジタル評価とは、コンピューターデバイスで実施される評価のことです。これには、評価対象の項目に応じて、さまざまに異なる形式の問いや刺激材料が含まれ、メディアをふんだんに取り入れたインタラクティブな評価から、従来の形式に近い小論文形式の試験まで、その種類は多岐にわたります。

実際のデジタル評価では、各生徒がそれぞれ個別のデバイスで試験に取り組むことになります。MYP で現在実施されているモデルでは、試験が事前にコンピューターデバイスに読み込まれるため、インターネット接続は推奨されているものの必須ではありません。デバイス起動中に、何らかの理由によってソフトウェアの動作が中断した場合は、試験ソフトウェアがデバイスを「ロック」し、それをすべて記録します（正当な理由でシステムがダウンした場合には、試験官が再起動できるようになっています）。MYP のデジタル試験は、オンライン上で実施することが強く推奨されています。これにより、一定の間隔で生徒の解答をオンラインで取得することができるため、学校側の事務処理の負担が軽減されます。MYP では、大多数の学校や生徒がオンライン機能を使っていますが、必ずしも試験の要件ではなく、IB ではオフラインでの受験もサポートしています。DP および CP のデジタル試験ではインターネット接続が必須とされ、生徒は、試験問題へのアクセスのみに制限されたロックダウンモードを使って試験に取り組みます。

出題形式が制限され、実施可能な範囲内の内容しか含むことができない紙の試験とは異なり、デジタル試験では、IB が本当に確認したい内容を出題できるという大きな利点があります。「デジタル評価が妥当性にもたらすメリット」のセクションで詳しく説明されて

いますが、デジタル評価の利点には、動画資料を含めることができる、小論文を執筆する際に生徒が共通のワープロツールを使用できる、生徒が色やフォントサイズを調整できるため受験上の配慮に対応できる、などが含まれます。

デジタル評価は今後も引き続き進化します。IB は、テクノロジーの進歩に合わせて評価をさらに開発していきます。現時点では、生徒がキーボードとマウスまたはトラックパッドのみを使用することを想定していますが、数年のうちにタッチスクリーンが主流となることで、さらなる可能性が広がると考えられます。さらに将来的には、仮想現実（VR）が評価に取り入れられるようになるかもしれません。VR 環境において、より実践に近い形式で外国語の評価を行うことは想像に難くなく、SF 映画などでは、技術的な課題（科学的な問い）を VR ラボで管理する未来も描かれています。

プログラム全体におけるデジタル変革に対する IB の継続的な取り組みによって、新たな用語が生み出され、より広い範囲において評価の関係者に影響を与えることになるでしょう。デジタル評価に関するする用語を含む IB 評価の用語集が、本資料の巻末に収載されています。

テクノロジーを使った指導

インタラクティブ機能をもつ教室のホワイドボードから、大人数授業を可能にする大規模なオープンオンラインコースまで、テクノロジーは生徒への指導に大きな変化をもたらしています。多くの生徒が、手書きではなくコンピューターデバイスを使って小論文を作成しています。

手書きからデジタル入力へと移行する流れは、評価と指導の実践の間に厄介な断絶を生みだしています。多くの生徒にとって 2〜3 時間にわたって手書きする経験は試験を受けるとき以外にほとんどなく、そのため、評価の正当性が著しく損なわれることになります。

デジタル評価の導入に伴い、生徒が評価の形式を理解していることを確認し、インターフェースの使用が学習内容への理解の実証の妨げとならないようにすることが大切です。これは練習ツールと模擬試験で達成することができますが、理想的な解決策は、指導においても同じツールを生徒に使用させることです。

優れたデジタル評価とは

デジタル評価は、従来の紙の試験では不可能な方法で試験問題を提起する手段としてテクノロジーを活用することで、評価の質を高めることができます。IB では、優れたデジタル評価を以下のように定義しています。

- テクノロジーを使って、妥当性と関連性が高く、現実に即した課題を設定する。
- 管理上の障壁をなくし、評価の実施を円滑にする。
- 評価の完了を妨げる新たな障壁をもたらさない。

e マーキング

「e マーキング」という用語は、IB が評価のために提出された成果物を採点する方法として、コンピューターデバイスを使用して答案を表示し、採点結果を記録する仕組みを表します。IB では、学習支援と多様な生徒の受け入れに関する例外的な状況を除き、試験官に試験や内部評価の成果物のハードコピーが送付されなくなりました。紙ベースの成果物はスキャンして電子形式にし、試験官が専用の採点ソフトウェアを使用してインターネット経由でアクセスできるようにしています。

e マーキングは、採点の速度と質に関して重要な利点を数多くもたらします。

- 評価のために提出された成果物は常に IB の管理下にあり、何らかの理由で 2 人目の試験官による評価が必要になった場合でも、成果物に即座にアクセスでき、国を越えて郵送する必要はありません。
- 成果物の匿名化により、採点における試験官のバイアスの可能性を低減することができます。
- より厳密な品質管理プロセスを実施し、採点中に採点基準を確認することができます。
- 1 つの学校の生徒が評価のために提出した課題を 1 人の試験官がすべて採点するのではなく、試験官全員で分担して採点できます。これにより、付加的な品質確認プロセスを実施することが可能になります。

e マーキングは 2010 年から IB で実施されており、生徒はデジタル形式で成果物を提出する必要はありません。現在、大部分の評価が手書きで実施され、それをスキャンしてデジタル形式に変換しています。

デジタル評価に関連したリスク

IB は、デジタル評価を開発するにあたり、生徒と学校にとって公平性を確保するために克服しなければならない重大な課題があることを認識しています。以下のセクションは、IB が管理している重要なリスクの一部のみを説明したものです。詳細については、本資料の個別のセクション、およびデジタル評価に関するその他の IB 資料に記載されています。

学校の負担

IB は、テクノロジーの使用が今日の生徒の日常に欠かせない要素となっており、評価においてもテクノロジーが組み込まれるべきだと認識しています。ただし、学校がプロセスを変更するには時間がかかり、デジタル試験には筆記試験とは異なる配慮が必要なことも認識しています。IB は、学校の所在地にかかわらず、学校に対して過度な期待を寄せることなく、テクノロジーの進化に合わせて DP と CP のデジタル試験をすべての学校に導入していく予定です。

デジタル評価への段階的な移行のサポート

IB は、より効果的で質の高い評価を実現するためにテクノロジーがもたらす機会を活用したいと考えるすべての学校がそれを実現できるよう、IB コミュニティーと協力しながら

時間枠を設定します。IB は、学習と教育の質、ひいては評価を向上させるために、テクノロジーの潜在能力を賢く活用すべきだと考えています。

テクノロジー障害

試験を受ける生徒にとっては、いかなるテクノロジー障害も許容されません。すべての生徒が、中断されることなく円滑に試験を受けられなければなりません。IB では、学校が事前に試験の練習をし、テストを実施し、規則が遵守されていることを確認するためのプロセスを用意しています。これまでの IB の経験によれば、ほとんどの問題は学校のインフラに起因して発生していますが、この包括的なプロセスを試験前に実施することにより、十分な時間的余裕をもって問題を特定し、解決することができます。

セキュリティー

デジタル試験では、少なくとも紙ベースの試験と同程度のセキュリティーを確保する必要があります。IB の評価は、試験の実施中は関連デバイスの他のすべての機能がロックされるように設計されています。MYP e アセスメントのように、試験前に評価資料が学校に送付される場合、これらのファイルは暗号化され、パスワードで保護されます。

IB は、関連業界におけるより高度なアプローチの開発に合わせてデジタル試験のセキュリティーを引き続き強化していきますが、同時に、現在の紙ベースのプロセスにも独自のセキュリティーリスクがあることも認識しています。

テクノロジーのためのテクノロジー

IB は、ただテクノロジーを導入したという実績をつくるためにテクノロジーを取り入れるのではなく、評価対象の質を向上するという目的のためにテクノロジーを使用します。IB は、テクノロジーの使用が評価の妥当性にどのような付加価値をもたらしたかを継続的に示し、これを評価開発プロセスに反映させていきます。

バイアス

IB は、試験に解答するための手段としてコンピューターデバイスを使用することで、成績が向上する生徒が一定数存在することを認識しています。ただし、紙ベースの試験にもバイアスが一切存在しないわけではありません。ペンで長時間書き続けるのが難しい生徒や、書くのが遅い生徒は、その時点で不利な立場に置かれています。

原則として、IB は、生徒の手書きスキルやタイピングスキルを測定する意図はなく、タイピングのスピードや書くスピードが速い生徒が有利にならないような評価課題を設計します。

IB は、こうした「デバイス効果」に関する研究論文に細心の注意を払い、この原則を満たすようにしています。

基準の上方修正または下方修正

IB の成績には意味があり、この意味は、IB の上級試験官による専門的な裁定を通して守られ、成果のデータによって裏づけられています。例えば、生徒が小論文に取り組む際、

段落をコピー・アンド・ペーストできた方が、構成の優れた小論文を作成しやすくなるということが考えられます。IB は、この点を考慮に入れて成績区分を設定するようにしています。これにより、紙ベースの試験とデジタル試験に同じくらい慣れている生徒が、どちらの試験においても、科目に対する理解と分析能力を反映した同等の成績を得られるようにします。

デジタル評価を通じて、IB が常に重視しながらもこれまで評価できなかった資質が新たにテストできるようになりました。これを受け、評価において生徒が実証することを期待されていた内容に調整を加える必要が生じる可能性があります。この点について詳しくは、「基準」のセクションを参照してください。ただしこの調整は、IB が長年守ってきた目標を維持し、現在の成績評価の説明に記載されている品質を反映できるような方法でのみ実行されます。

関連文献

IB のデジタル評価についての詳しい情報は、以下のリソースをご覧ください。

- プログラム・リソース・センターの「Transition to Digital Examinations（デジタル試験への移行）」のページ
- 「MYP eAssessment playbook（MYP e アセスメントの手引き）」
- 「MYP eAssessment: Introduction for school leaders and teachers（MYP e アセスメント：学校リーダーおよび教師向けの概要）」（動画）
- 「MYP eAssessment Ready! Introduction（MYP e アセスメントの開始に向けて）」

教育における評価

- 教育において、評価はしばしば形成的評価（学習**のための**評価）と総括的評価（学習**の**評価）に分けられてきました。現在は、「学習**としての**評価」を目指す動きが強まっています。
- 特定の文脈において使われる評価の種類は、その成果の用途、または評価の目的によって決定される必要があります。
- 評価は、指導の実践に影響を与える可能性があり、「逆流効果」（評価が学習と指導に及ぼす影響）が有益なものになるように設計されていなければなりません。

評価を定義する

IB 資料『国際バカロレア（IB）の教育とは』（2019 年発行）に記載されているように、IB は生徒に積極的な役割を提供し、効果的な学習に対する文脈の重要性も認識する、学習への構成主義的なアプローチを採用しています（Murphy, 1999）。評価を使って効果的な学習と指導を支えるには、この構造主義的な学習理論を軸にして評価を設計する必要があります。この概念に関連した研究について詳しくは、参考文献（Shepard（1992）、Wood（1998）、Black（1999、Lambert & Lines（2000））を参照してください。

「評価」は、生徒の成果を収集して評価するあらゆる方法を意味することができます。一般的な評価として、テスト、試験、実習、プロジェクト、ポートフォリオ、口述課題などが挙げられます。評価は長期間にわたって実行される場合もあれば、数時間で完了する場合もあります。評価は、その種類と目的に応じて、教師、他の生徒、生徒自身、または試験官によって採点されます。

図 5
さまざまな種類の評価の例

生徒の資質（能力）をテストする評価課題では、**補完**モデルまたは**習熟**モデルのいずれかを使用できます（図 6 を参照）。

ほとんどの外部試験では**補完**モデルが使用されており、部分的に成績が低くても、他の部分で高い成績をとることでその分を補うことができます。

一方、**習熟**モデルでは、評価の各部分に、達成しなければならない最低レベル（習熟度）が設定されています。

3 つの設問から構成され、各設問が 10 点満点の試験があります。

合格点は 15 点ですが、これは各設問で 5 点ずつ獲得して 15 点にしてもよく、10 点、3 点、2 点で 15 点にしてもかまいません。後者の場合、最初の設問で満点をとることで、他の 2 つの設問で点数を失った分を補っています。

同じ試験の例（3 つの設問から構成され、各設問が 10 点満点）で考えてみると、各設問の合格点が 5 点に設定されている場合、最初の生徒は合格ですが、2 番目の生徒は不合格となります。10 点、10 点、4 点をとった生徒も不合格となります。

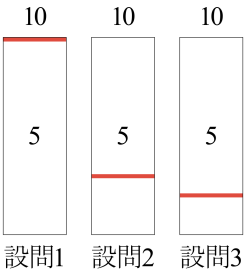
図 6
補完モデルと習熟モデルの比較

補完モデル

設問数：3
評点数（各設問）：10
合格最低点（合計）：15/30

補完モデルを適用：

全設問の合計点が合格最低点以上であるため、評価は**合格**

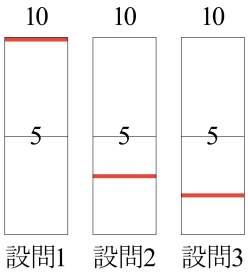


習熟モデル

設問数：3
評点数（各設問）：10
合格最低点（設問あたり）：5/10

習熟モデルを適用：

設問2および3の評点が足りないため、評価は**不合格**



習熟モデルの評価は、キャリア関連または職業関連の試験で多く使われます。このような分野では、1 つのことには秀でているが他のことは苦手であるという状況は望ましくありません。例えば、洋服を縫いあげるときに、基本的な裁縫スキルが欠けている場合、たとえデザインスキルにどれほど優れていようともその部分を補うことはできません。

IB では、さまざまな種類の評価ツールを活用しています。これには、プログラムの終盤に受験する試験に加え、それぞれの科目のコースのさまざまな時点において、多様な条件下で実施される多種多様な課題（研究論文、記述課題、口頭試問、科学研究や数学研究、フィールドワークプロジェクト、芸術パフォーマンスなど）が含まれます。

評価のアプローチ

評価はさまざまな目的に使うことができます。評価の目的は、その評価の設計方法に大きな影響を与えます。従来、評価には形成的評価と総括的評価という 2 つの大まかな種類がありました。

形成的評価は、生徒の学習方法に直接つながるもので、「学習**のための**評価」と呼ばれることがあります。一方、総括的評価は「学習**の**評価」と呼ばれることがありますが、これは、総括的評価が教室での実際の学習内容に及ぼす大きな影響を過小評価しています。評価はすべて、適切な学習を支えるものであるべきです。総括的評価は、単に学習が行われた後に実施される活動としてではなく、学習と指導において統合的な役割を果たすように設計される必要があります。

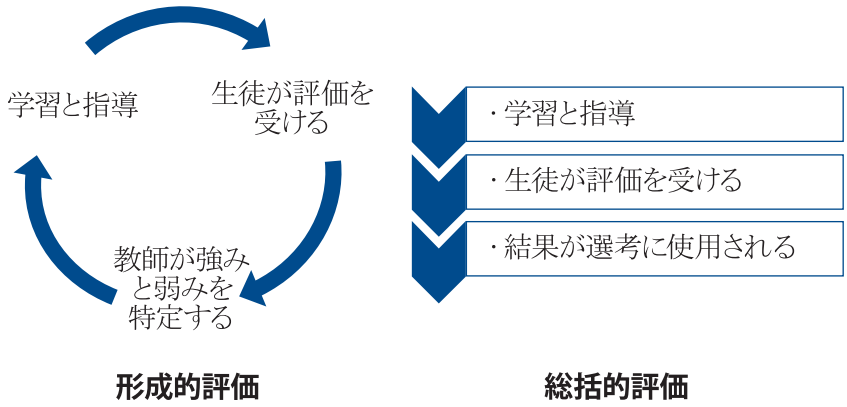
形成的評価のねらいは、生徒の強みと弱みの本質について詳細なフィードバックを教師と生徒に提供し、生徒の能力の発展をサポートすることにあります。ここでは、教師と生徒の間のディスカッション形式といった直接的なやりとりを含む、特定の評価方法が特に役立ちます。

ヴィゴツキー（Vygotsky, 1962）は、教師は学習の先導者ではなく、学習のサポートを行う者と捉えられ、評価課題や評価ツールを駆使して「最近接発達領域」における学習をサポートすべきであると述べています。これは、生徒が自分自身で達成できるものと、教師のサポートがあって達成できるものの間に存在する達成領域のことを指しています。

「スキャフォールディング（足場づくり）」という概念は、ウッドら（Wood et al., 1976）によって提唱されました。この枠組みでは、教師は学習構築のための足場を提供しますが、実際に学習を構築できるのは生徒のみです。教師は、生徒にとって最適な難易度の形成的評価を設定し、生徒の進捗に合わせて難易度を調整し続けることを目指す必要があります。

その一方で総括的評価は、生徒が**できること**を測定することに重点を置いています。通常、トレーニングプログラムを修了したことを実証したり、教育の次の段階に進む準備ができていることを示したりするために使用されます。形成的評価では、生徒の取り組みの背景にある**理由**を探ろうとしますが、総括的評価では、その取り組みが**正しい**かどうかを把握しようとします。理由を探る方が有益なように思えるかもしれませんが、総括的評価の目的は生徒について判断を下すことであり、将来の指導に役立つ情報を得ることではありません。「総括的評価における生徒の成果の説明」のセクションで、総括的評価の必要性についてさらに詳しく説明しています。

図 7
形成的評価と総括的評価の活用方法



形成的評価では、各生徒の到達度を正確に測定することよりも、生徒がまだ身につけていない知識、スキル、理解を正しく特定することが重要です。これらの優先順位の間でバランスをとることを「妥当性」と呼び、後のセクションで詳しく説明します。生徒の到達度とフィードバックの質の間のバランスは、総括的評価では逆転します。総括的評価では、評価の結果が、就職や進学における選考プロセスなどにおいて生徒についての判断を下すために使用されますが、今後の指導をサポートするためにも使用されることもあります。

さまざまな国の評価システムを分析すると、多岐にわたる評価手法やアプローチが使われていることが見てとれます。すべてのシステムが、テクノロジー、リソース、時間をめぐる考慮事項、および国の教育システムに及ぼす影響という点で、それぞれに強みと弱みを有しています。たとえ、特定の文脈において、まったく新たな評価システムを一から設

定することができる場合でも、あらゆる状況にあてはまる最適な評価実践というものは存在しません。評価システムを考案する際に行われる選択は、必然的に、その評価システムが属する広範な社会的文脈がもつ価値観と優先順位を反映することになります。この分野における研究については、クレスウェル（Cresswell, 1996）とブロードフット（Broadfoot, 1996）の論文を参照してください。

また、重要な点として、総括的評価が教育の質を測る尺度として使われるケースが増えてきているということが挙げられます。これにより、評価を実施する理由について、生徒のためではなく教育システムのためという新たな側面が加わります。

逆流効果と学習

> 必要とされているのは、できる限り妥当性の高い評価プロセスであり、その逆流効果によって質の高い指導を妨げたり、適切なタイミングを逃したりしてはならず、ただでさえ不足している教育的リソースを過度に使用するものであってもなりません。
>
> （Peterson, 1971）

この引用において、IB の創設者であるアレク・ピーターソンは、評価の実施方法が学習へのアプローチに影響を及ぼす可能性があるというリスクを認識しています。サージェナー（Surgenor, 2010）は、評価は生徒の学習と指導の経験に良い影響と悪い影響の両方を与える可能性があると述べています。

スナイダーは、1971 年の論文で、評価において何が重要になるかという暗黙的および明示的なメッセージに基づいて、生徒がカリキュラムに対する独自の理解を形成することを提案しました。「隠れたカリキュラム」と呼ばれるこのアプローチは、科目で合格点をとる方法は理解しているのに、科目そのものは理解できていない状況につながる可能性があります。これを理解するもう 1 つの方法として、ギブスの以下の引用が挙げられます。

> 試験を再受験した時、単に試験に合格することだけに集中した。再試験では 96%を獲得し、なぜ 1 度目で合格できなかったのかといぶかしがられた。だから、今回は科目を理解することよりも、試験に合格することに注力したんだと伝えた。いまだに科目については理解できていないので、ある意味で本末転倒だといえる。
>
> （Gibbs, 1992, p. 101）

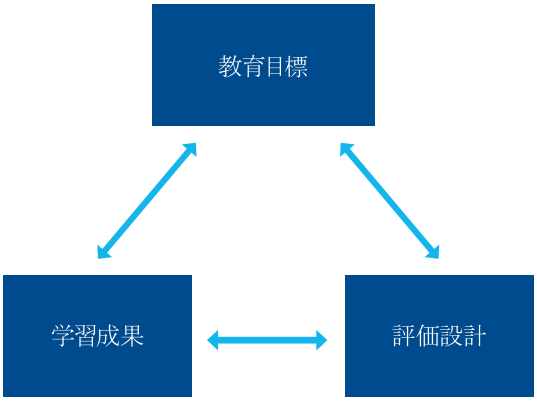
この逆流の概念は、「テストに出ないことは指導されない」という格言にも表現されています。これは、教科と指導は切り離して考えることができないという点を強調するものです。

評価は、生徒に教育目標を実証するよう促すものですが、評価が教育目標をサポートできない理由としてさまざまなものが考えられます。最も可能性が高いのは、評価の根拠となる学習成果が教育本来の目的を効果的に反映していないことです。例えば、トップアスリートを養成する職業訓練コースでは、陸上競技のルールを理解することが学習成果の中核となっている場合があります。もう 1 つのよくある問題は、評価では意図されたすべての成果が網羅されているものの、実際のテストのほとんどが、全体的な教育目標にとって特に重要な課題ではなく、テストすることが容易な課題に重点を置いている可能性があることです。

このような評価と教育目的の相互依存性は、図 8 に表現されています。これは、ファーストのパラダイム（Furst, 1958）をフリスとマッキントッシュ（Frith & Macintosh,

1984）が改変したものです。ここに示す 3 つのうち、いずれかの 2 つの要素の間のつながりが欠落している場合、ほぼ確実に評価の質が低下します。

図 8
目標、成果、設計の間の関係性



妥当性

- 「妥当性」という言葉は、評価が**目的に適っているかどうか**を示しています。これは多くの側面をもつ複雑な概念です。
- 評価は多くの目的で使われます。
- ある評価が、1 つの目的にとっては妥当であっても他の目的には適していないということもあり得ます。例えば、綴りのテストで言語の流暢さを測定することはできません。
- 妥当か妥当でないかは、評価自体ではなく、評価の結果が使用される目的に応じて判断されると論じることもできます。
- IB ではまず、プログラムの妥当性を優先事項とし、その次にプログラムの要素（個別のコースなど）が妥当かどうか、そして最後に個々の評価が妥当かどうかを重視します。

妥当性を定義する

試験を受ける目的は何か。何のための試験か。これらの問いは、評価が「目的に適っている」または「正しいことを評価する」とは何を意味しているのかを理解するにあたり、大きな重要性をもちます。多くの場合、これらの問いに対するさまざまな答えは互いに拮抗することになります。また、教育の目的のすべてを特定の試験でテストできるとは限らないということも考えられます。実際、目的によっては、テストすることが不可能なこともあります。

例えば、「数学」の評価の目的に、以下の 4 つが含まれるとします。

1. 学習プログラムを終えた生徒が何を理解しているのかを認識する
2. 進学や就職における選考手段として機能する
3. 将来の成功に関する指標を提供する
4. プログラムのカリキュラム目標の指導を強化する

これらの単純な目標同士であっても、適切なバランスを見つけることは簡単ではありません。幾何学よりも微積分が、「数学」における将来的な成功に貢献する度合いが大きい場合、評価は微積分に重点を置くべきでしょうか。これは、（幾何学において）生徒が何を理解しているのかを認識するという最初の目的にどう関連するでしょうか。評価を設計するにあたって、2 つ以上の目的が相反した場合はどうなるのでしょうか。

上記の 4 つの目的は、すべてを網羅したものではなく、規定的なものでもありません。ニュートン（Newton, 2007）は、評価結果の活用方法について、多くの例を示しています。

・ 社会的な評価としての使用 ・ 形成的な使用 ・ 生徒のモニタリングのための使用 ・ 転移のための使用 ・ クラス分け（レベル分け）のための使用 ・ 診断のための使用 ・ ガイダンスとしての使用 ・ 資格認定のための使用 ・ 選考のための使用	・ ライセンス付与のための使用 ・ 学校の選択のための使用 ・ 機関のモニタリングのための使用 ・ リソースの配分のための使用 ・ 組織的な介入のための使用 ・ プログラム評価のための使用 ・ システムのモニタリングのための使用 ・ 同等性のための使用

このように評価の使用方法は互いに対立していますが、個人および機関はそれでも、幅広い文脈において、人に関する決断を下さなければなりません。評価結果が利用できない場合は、その他の方法として、それほど効果的に設計されておらず、バイアスを含む可能性のある手段を使ってこの決断を下すことになります。この問題点について、クレスウェル（Cresswell, 1986）は以下のように述べています。

> 他の規準の信頼性が試験よりも低い場合、それらへの依存を高めると、選考をめぐる意思決定の信頼性が低くなることは明らかである。
>
> （Cresswell, 1986, p. 42）

資格認定（または資格認定のための個別の評価など）が「意図された目的に適合している」という概念は、広く言えば妥当性を意味します。

妥当性は評価の質に関連する重要な用語としてよく使用されますが、学術界では長きにわたり、その意味を正確に定義しようとしてきました（Newton、2012）。さらに複雑なことに、しばしば**信頼性**と併せて使用される、より狭い概念としての妥当性も存在します。これは、評価が測定するとされているものを実際にどの程度測定しているかを表します。混乱を避けるため、本資料では 2 つ目の概念を「構成の関連性」と呼びます。「構成の関連性と現実に即した評価」のセクションを参照してください。

特定の評価が妥当かどうかの決定は、最終的には個人の判断に基づきます。この判断は、テストの使用目的に応じたテストスコアの解釈を支える体系的なエビデンス群に基づいて下されます。このセクションの前半部分で特定された拮抗する評価のニーズは、後ほどさらに詳しく説明しますが、これらのニーズの間には常に妥協点が存在します。重要な問いは、利用可能なエビデンスが、その評価が十分に目的に適い、有用であることを示唆しているかどうかです。したがって、妥当性について議論する際は、単純に評価が妥当である、または妥当ではないと宣言するのではなく、このエビデンスの説得力（妥当性に関する議論の説得力）について議論することが優れた実践といえます。この理由により、妥当性の是非についての判断は、エビデンスに基づく判断となります。妥当性は、白か黒かの二項対立ではなく、文脈的であり目的を基盤とします。ニュートンとショーは、妥当性について主張する前に、評価の目的への適合性を考慮する必要があり、「妥当性の確認は、テストの点数の解釈および使用に関する議論の説得力を高め、評価するプロセスと見なさ

れる」と説明しています（Newton and Shaw, 2014, p. 3）。したがって、妥当性は評価自体がもつ性質ではなく、評価結果から導かれる推論の性質といえます。

最後に、妥当性は、評価の設計においてのみ達成されるものではありません。むしろ、評価のライフサイクル全体を通して、継続的に発展していくものです。同様に、妥当性の議論はライフサイクルの最初に行われるものではなく、評価の設計プロセスを通して継続的に追加され、磨き上げられていきます。

妥当性の議論を構築する

妥当性は、単純な目標を示す概念ではなく、対立する問題の妥協点を見出すことを意味します。妥当性を「証明」することはできません。できるのは、コースや評価に関して下された意思決定が、そのコースや評価を有意義なものにするとともに、意図された目的に適ったものにする理由について、説得力のある議論を構成することのみです。つまり、妥当性をもつのは評価（または評価結果）ではなく、それが適えようとする目的ということになります。ニュートン（Newton, 2012）は、「ある特定の目的（つまり、ある特定の決定を下すため）に所定の評価手順を使用することは、その解釈をめぐる議論に十分な説得力がある場合に妥当である」と述べています。

同様の理由から、妥当性の議論のエビデンスは、プログラムを開発し、モニタリングし、実施する過程で自然に生じるべきであり、この過程において行われた議論や下された決定を反映する必要があります。これは、例えば評価において、特定の試験でどの問題を出題するかの判断が、カリキュラムが適切に含まれることを示すものとなっていることを意味します。

重要と考えられるすべての側面をカバーする適切な構造があることは、妥当性をめぐるあらゆる議論に不可欠な要素です。IB においてこれは、エビデンス収集の対象となる一連の問いによって表されます。

妥当性を維持する

妥当性は、評価モデルの設計時にのみ決定されたり「証明」されたりするものではありません。むしろ、評価のライフサイクルを通じて、さらに言えば、評価の結果に基づいて決定が下される限りにおいて、評価とともに発展し続けます。

IB にとって、ある評価が妥当であることを示すさまざまなエビデンス（つまり妥当性に関するさまざまな議論）は、特定のコースの期間全体にまたがる多くの場面で収集されるものです。

コースの開発時や改訂時に、コースの目的、およびその目的を満たすための評価方法をめぐる議論がなされ、それが妥当性に関する議論の中核を成します。特に、構成の関連性と、妥当性のその他の側面とのバランスが検討されます。すべてのセッションについて、新しい試験や評価課題を作成することは、特に信頼性、公平性、同等性、管理のしやすさ、構成の関連性に関してより多くのエビデンスを生み出します。これにより、教師からのフィードバックと生徒の試験結果は、評価がその目的をどの程度達成できたかを示す指

標となり、採点の信頼性と成績の同等性に関するエビデンスを提供します。そしてこの情報は、次の評価の開発にフィードバックされます。

最後に、特定のコースのすべての評価と、特定のプログラム内のすべての学習コースから収集されたエビデンスは、次にコースやプログラムが見直される際に意思決定の根拠となります。

> テストの妥当性を評価することは、一度限りの静的な取り組みではなく、継続的なプロセスです。
>
> (Sireci, 2007, p. 477)

妥当性の鎖の構成要素

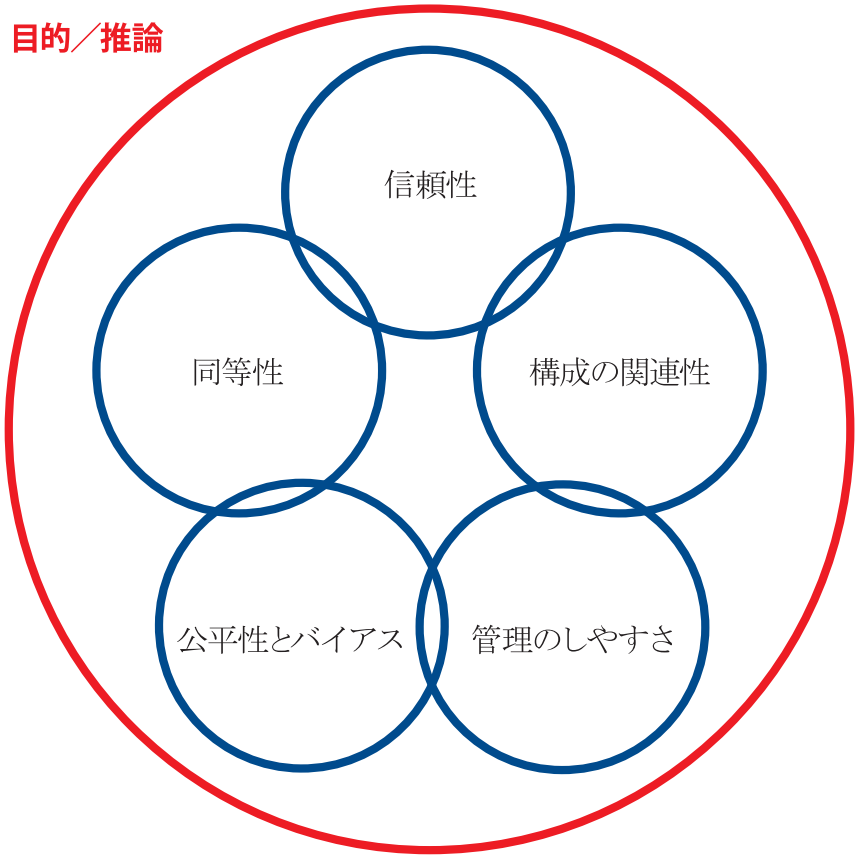
- 妥当性には多くの異なる要素があります。これらの要素は鎖のようにつながっていて、1 つの環が壊れると、評価全体の有効性が損なわれることになります。
- IB では妥当性の要素として、信頼性、構成の関連性と現実に即した評価、管理のしやすさ、公平性とバイアス、および同等性に焦点をあてています。
- この 5 つの要素はしばしば、互いに対立します。評価が妥当かどうかを検討する際は、その評価の主な目的について考え、各要素の相対的な重要性を決定する必要があります。
- IB は、構成の関連性を実証するプログラム、コース、評価の作成に最も重点を置いています。
- 評価に関して言えば、採点の信頼性を確保しやすい試験ではなく、より高度な思考スキルをテストする有意義な課題や設問を提供することに重点を置いています。

妥当性と信頼性は、あらゆる評価システムの重要な特性だと見なされています。これは特に、評価の結果が生徒や教師に大きな影響を及ぼす、重要性の高い評価にあてはまります。これらの特性は多面的であり、さまざまな種類の妥当性と信頼性が存在します。

妥当性の複雑さは、**妥当性の鎖**という考え方で表現することができます（Crooks et al., 1996）。鎖を構成する 5 つの要素はそれぞれに重要性をもちますが、いずれの要素も、単独で評価の妥当性を確保する（つまり目的に適合する）ことはできません。例えば、信頼性が非常に高い評価であっても、体系的に特定の集団に不利に働く、ということが考えられます。あるいは、意図された内容に正確に焦点をあてた課題であっても、その長さと要件により、生徒のスタミナや学校の対応能力が試されることになる場合があります。次のセクションでは、これらの各要素についてさらに詳しく説明します。妥当性を実現するにはすべての要素が必要ですが、各要素の間には対立関係も存在します。例えば、試験においてカリキュラムのすべての側面を網羅しようとすると、生徒にとって試験時間が長すぎるものとなってしまうでしょう。同様に、テストが実施される特定の文化や国に合わせて内容を文脈化しようとすると（公平性）、それぞれのケースで本当に同じテストが実施されるのか（同等性）という疑問が生じることになります。

評価が妥当性をもつためには、鎖のすべての環が欠けることなく存在していなければなりません。

図 9
妥当性の鎖



妥当性の各側面のバランスをとる

評価を設計する際には、評価の目的に照らして妥当性の相反する要素のバランスをとることが重要です。単一の評価で各要素の最高水準を達成することは不可能であるため、妥協点を見つける必要があります。同様に、一部の要素については評価の設計時に決定され、変わることはありませんが、その他の要素（特に信頼性と公平性）は、評価が実施され、採点と成績付与が行われる過程で変化していきます。

図 10
妥当性の要素の間で優先順位のバランスをとる：公平性と管理のしやすさ

		
バランスが悪い： 受験者に過度な負担がかかっている（管理のしやすさ）	**バランスが適正：** 受験者が妥当な方法で評価される	**バランスが悪い：** 出題されるトピックが運に左右される（公平性）
評価が50問の論述問題（各20分）から構成され、コースのすべての要素を取り扱っている。 つまり、評価の所要時間が16時間となる。	評価が20問の短答式問題（各2分）、5問の論述問題（各10分）、2問の発展課題（各30分）から構成される。各設問がコースの異なる要素を取り扱っている。 評価の所要時間が合計2時間半で、コースで学習したトピックの半分強を取り扱っている。	評価が、コースで学習した50の要素のうち1つの要素についての知識と理解を問う30分の発展課題1問のみで構成されている。 受験者の最終成績が1つの課題のみで決まる。

妥当性の 5 つの要素は個別の定義をもちますが、それぞれがどのように顕在するか、また、それぞれをどのように管理するかという点で、大きく重複しています。5 つすべてを合わせることで、IB 評価における妥当性の広範な概念が形成されます。

最後に、IB は**構成の妥当性**、つまり評価の内容が、評価対象である資質や能力を確認するものとなっているかどうかを最も重視しています。それでも、この要素を優先するあまりに、妥当性の鎖を構成する他の要素を犠牲にすべきではないということを覚えておくことが重要です。

信頼性

評価における信頼性とは、同じテスト手順を繰り返した場合に、生徒のテストの結果が前回と同じものになる度合いとして定義されます。これは必ずしも、生徒が「正しい」結果を得ることと同義ではありません。

リンクリーとクレスウェル（Winkley & Cresswell, 2011）は、信頼性の概念を紹介するにあたって、信頼性に対する潜在的なリスクの一覧を提示するとともに、信頼性を確保することが望ましい領域（生徒のパフォーマンスの採点など）を列挙しています。

1. **採点者間の信頼性**：特定の設問について、ある試験官の採点が他の試験官の採点よりも厳しくなる、または甘くなるということが考えられます。同様に、同じ試験官でも、日によって採点が甘くなったり厳しくなったりすることもあり得ます。
2. **生徒のパフォーマンスの変動性**：試験における生徒の成績は、日によって変わる可能性があります。これは特に、試験の実施状況が変わる場合にあてはまります。試験の実施状況として、試験の実施時間が午前か午後か、試験監督は誰か、前日の夜によく眠れたか、外で誰かが芝刈りをしていないか、などが含まれます。

3. **異なる試験問題**：出題される問題が試験ごとに異なり、これによって生徒の理解のさまざまな側面がテストされる可能性があります（通常、テストはカリキュラムから抜粋されます。これは、すべてをテストするには時間が足りず、生徒は 1 つのトピックにしか取り組まない場合があるためです）。
4. **年ごとに比較した場合の成績の同等性**：この同等性を確保することで、生徒の成績を年ごとに有意に比較できるようになります。
5. **試験の設定の違い**：試験の設定やシラバスに変化があった場合、経時的な同等性を確保することが難しくなります。
6. **さまざまな種類の評価活動**：多くの資格認定が、さまざまな種類の評価活動から構成されています。このような評価方法の違いによって、評価の信頼性に関してさまざまに異なる課題がもたらされます。
7. **さまざまな種類の設問**：出題される設問の種類によって、生徒の成績が変わることがあります。

実践の観点から、IB は通常、生徒が評価を一度受けると想定し、プロセスの結果に**一貫性**を見出そうとします。そのため、IB は上記リストの 1、3、6、7 に焦点をあてることになります。

資格授与機関である IB は、評価の信頼性を高めるための措置を講じています。採点の信頼性はこの中心的な側面であり、標準化、品質モデル、モデレーションは、主に試験官の採点基準に一貫性をもたせることを目的としています。

採点における信頼性

次の簡単な演習は、採点における信頼性の概念と、それが実際にどのように機能するかを示すものです。

図 11 は、IB の「言語の習得」の評価における実際の答案から抜粋した、5 つの成果物を示しています。

マークスキームを使って言語の質を 8 点満点で評価し、それぞれの成果物に付与したいと思う評点を記録します。

各成果物の採点が終わったら、採点の質に関するフィードバックを確認します。フィードバックは生徒の答案の後にあります。

マークスキーム

生徒は振り返りスキルをどの程度効果的かつ正確に適用したか。

評点	レベルの説明
0	生徒は、以下に記されたレベルのいずれの基準にも達していない。
1〜2	プロジェクトが自分の学習に及ぼした影響、およびねらいが達成されたかどうかを述べている。
3〜4	プロジェクトが自身の学習に及ぼした影響、およびねらいが達成されたかどうかを簡単に説明し、それがエビデンスによって部分的に裏づけられている。

評点	レベルの説明
3〜4	プロジェクトが自身の学習に及ぼした影響を説明し、成功規準に基づいて成果を評価し、それがエビデンスによって裏づけられている。
7〜8	プロジェクトが自身の学習に及ぼした影響を詳しく説明し、成功規準に基づいて成果を評価し、それが具体的なエビデンスまたは詳細な説明（もしくはその両方）によって完全に裏づけられている。

順番に答案を確認し、手引きとしてマークスキームを活用しながら各答案に評点をつけます。

図 11A
採点の信頼性の演習

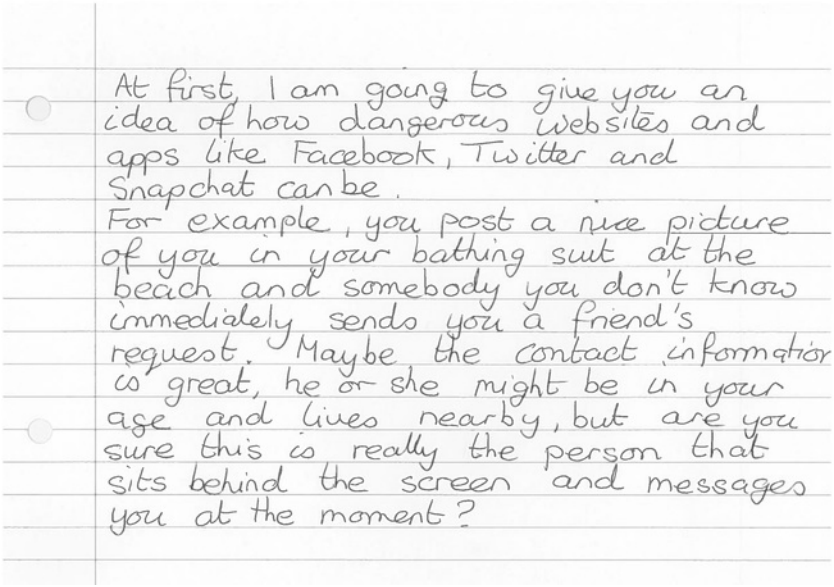

At first I am going to give you an idea of how dangerous websites and apps like Facebook, Twitter and Snapchat can be.
For example, you post a nice picture of you in your bathing suit at the beach and somebody you don't know immediately sends you a friend's request. Maybe the contact information is great, he or she might be in your age and lives nearby, but are you sure this is really the person that sits behind the screen and messages you at the moment?

図 11B
採点の信頼性の演習

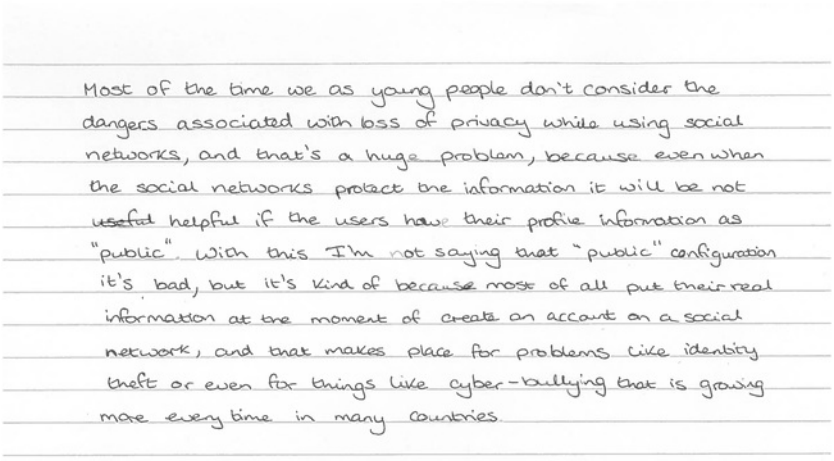
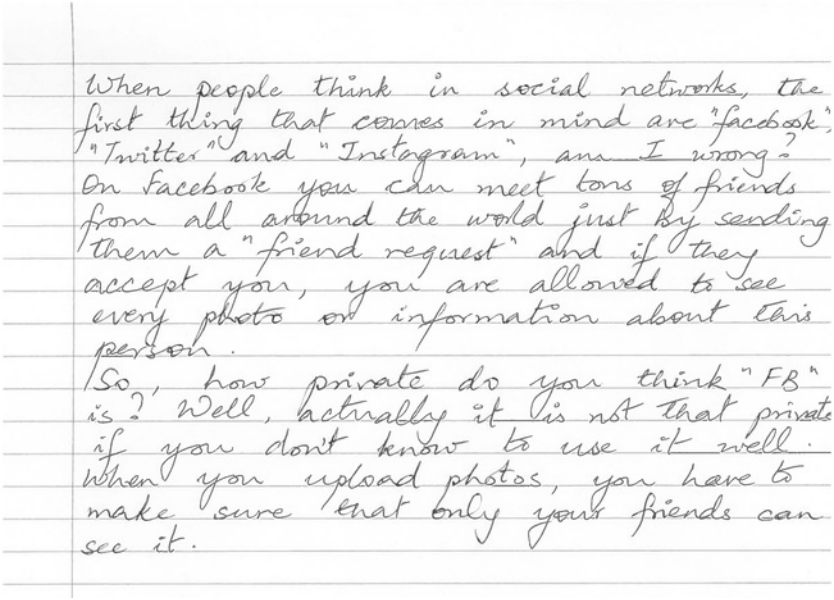
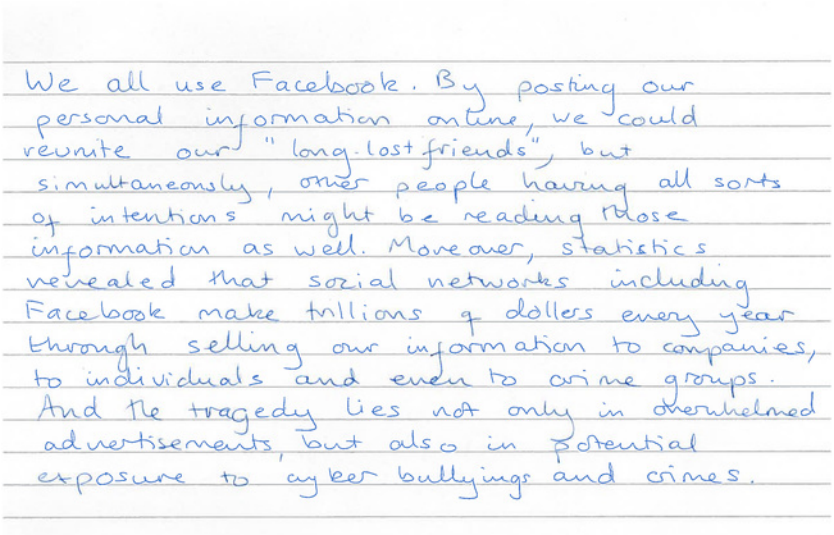

Most of the time we as young people don't consider the dangers associated with loss of privacy while using social networks, and that's a huge problem, because even when the social networks protect the information it will be not useful helpful if the users have their profile information as "public". With this I'm not saying that "public" configuration it's bad, but it's kind of because most of all put their real information at the moment of create an account on a social network, and that makes place for problems like identity theft or even for things like cyber-bullying that is growing more every time in many countries.

図 11C
採点の信頼性の演習



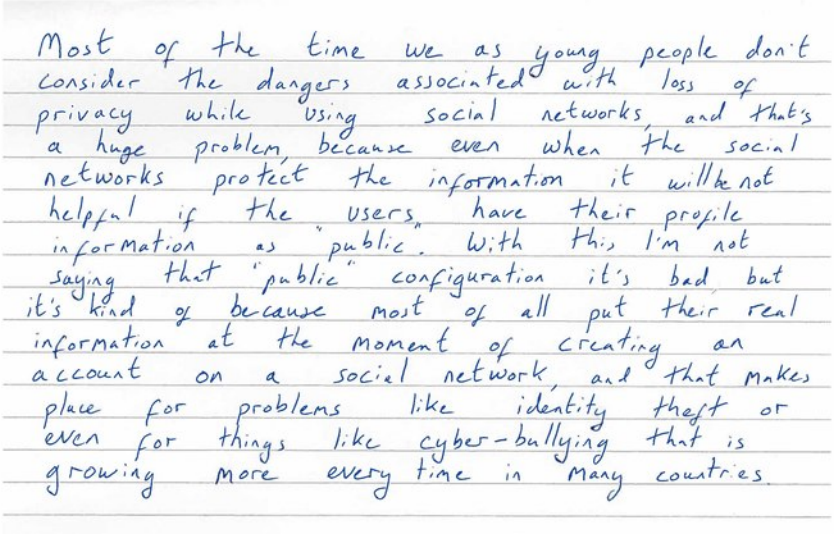
When people think in social networks, the first thing that comes in mind are "facebook", "Twitter" and "Instagram", ana I wrong? On facebook you can meet tons of friends from all around the world just by sending them a "friend request" and if they accept you, you are allowed to see every photo or information about this person.
So, how private do your think "FB" is? Well, actually it is not that private if you don't know to use it well. When you upload photos, you have to make sure that only your friends can see it.

図 11D
採点の信頼性の演習



We all use Facebook. By posting our personal information online, we could reunite our "long-lost friends", but simultaneously, other people having all sorts of intentions might be reading those information as well. Moreover, statistics revealed that social networks including Facebook make trillions of dollers every year through selling our information to companies, to individuals and even to crime groups. And the tragedy lies not only in overwhelmed advertisements, but also in potential exposure to cyber bullyings and crimes.

図 11E
採点の信頼性の演習


Most of the time we as young people don't consider the dangers associated with loss of privacy while using social networks, and that's a huge problem, because even when the social networks protect the information it will be not helpful if the "users" have their profile information as "public". With this I'm not saying that "public" configuration it's bad but it's kind of because most of all put their real information at the moment of creating an account on a social network, and that makes place for problems like identity theft or even for things like cyber-bullying that is growing more every time in many countries.

各答案の採点を終え、各生徒に評点を付与したら、採点に関するフィードバックを確認します。

フィードバック

生徒 B と生徒 E は、筆跡は異なるものの同じ成果物であることに気づきましたか。生徒 B と生徒 E に同じ評点を与えましたか。

特定のマークバンド（採点基準表）内の 2 つの評点のうち、どちらを与えるか迷ったかもしれません。もう一度採点をする場合、同じ判断を下す自信はどれくらいありますか。

ここでのポイントは生徒の成果物の質ではなく、同じ人が成果物を採点するたびに、前回とは若干異なる判断を下すことがあるという点です。大まかに言えば、成果物の質に対する見方は毎回同じですが、実際に付与する評点が若干異なる可能性があるということです。

数日後に各課題をもう一度採点し、2 回目の評点が今回とどのように異なるかを確認します（今回の評点はなるべく忘れるようにしましょう）。

IB が採点の信頼性を確保する方法について詳しくは、「Reliability: How we make sure marking is fair（信頼性：採点の公平性を確保する方法）」（動画）を参照してください。

一貫性のある成果と「正しい」成果

高い信頼性を確保するうえで目指すべきは、どの試験官が成果物を採点した場合でも、「正しい」評点ではなく、同じ（公平な）評点が付与されるようにすることです。生徒の成果物の良し悪しは専門的判断に委ねられ、2 人の教師の間でどの評点を付与すべきか意見が分かれることもよくあります。信頼性で重要なのは、2 人の教師が（上級試験官と）同じ判断を下すことです。

これは、成績照会サービスへの対応において、特に難しい問題となります。このような場合、試験官は、これが初めての採点のつもりで評点を付与し、成績区分や個別の生徒に

与える影響といった現在わかっている追加情報によって、採点結果が（肯定的にも否定的にも）左右されないようにする必要があります。

教育以外の分野において評価の信頼性に対する理解が乏しいことはよく知られていますが、試験結果に関する公の議論が増えるのに伴い、このトピックの重要性をより強調する必要があります。

> 研究論文などで示唆されているように、評価の信頼性、特に測定の不正確さは、参加者にとって理解が難しい概念とされています。
>
> (Chamberlain, 2010, p. 3)

> 個人の能力を測るうえで、試験は公平性のある測定方法ではないのではと考えることがあります。試験は、その時点での個人の状態と、その個人や他の受験者がどの程度勉強したかに焦点があてられています。試験で高い点数をとれない人間が、後になって頭角を現し、分野の第一人者となる可能性もあります。(健康分野で働く男性)
>
> (Chamberlain, 2010, p. 27)

> 教師と生徒の約 63%が「いかなるレベルの間違いも許容されない。たとえ 1 人でも生徒が間違った成績を付与されるということは、まったく許容されない」を選択した一方で、50%以上が「試験での誤記のようなケアレスミスと、2 人の採点者の採点が一致しないという避けられない不一致の間には違いがある」を選択し、このような不一致に対する寛容さを示唆しています。この矛盾は、信頼性に関する知識の有無と、信頼性の欠如に対する態度との相関が希薄であることを、示している可能性があります。
>
> (He et al., 2010, p. 27)

構成の関連性と現実に即した評価

私たちは、測定しようとしているものをどの程度正確に測定しているのでしょうか。測定の対象とされるスキルや知識を正確に測定することは、「構成の妥当性」と呼ばれることがあります。ただし、大局的な意味での妥当性との混同を防ぐため、本資料では**構成の関連性**という言葉を使います。

現実に即した評価という考え方は、構成の関連性と密接に関係しています。**現実に即した**という言葉は、実用性を損なわない限りにおいて、生徒が実社会で遭遇するであろう状況に合致した方法で試験を実施していることを意味します。現実に即してない評価の例として、文脈と課題を切り離した評価、過度に簡略化された評価、明らかに不自然な評価などが挙げられます。

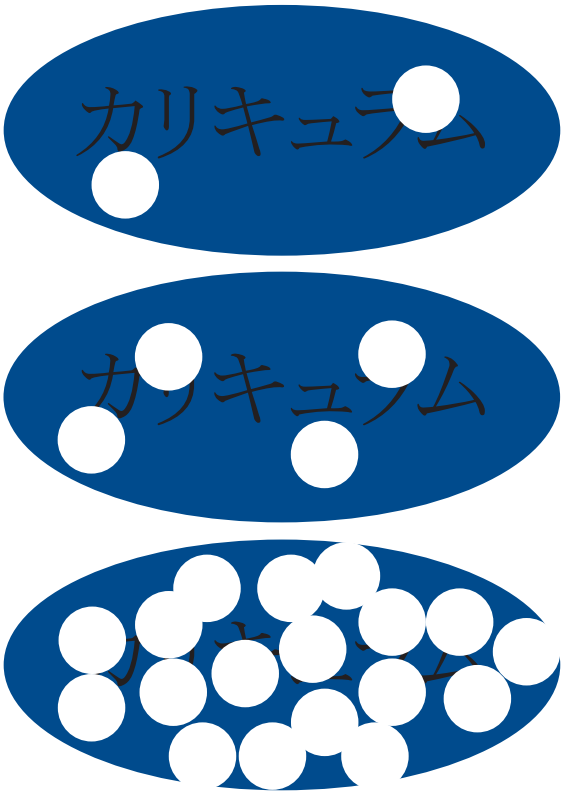
構成の関連性が低い、拙悪な評価設計の例を挙げることは難しいことではありません。例えば、口述試験で手紙を書くスキルをテストすることがこれにあたります。ただし、課題設定の方法として広く受け入れられているアプローチが、目的の内容を本当の意味でテストしていないというケースも多くあります。例として、従来の文学試験の構成を考えてみましょう。生徒は、所定のテクストの中で使われている文学技巧について学習した後、評価課題として、そのうちの 1 つについて小論文を書くように求められたとします。仮にこの生徒が、教師が言ったことをすべてそのまま覚えていたとしたら、この課題は、文学に対する理解を実証することになるのでしょうか。

特定の評価課題が生徒に実証させようとしていることを明確に理解し、それがどの程度実証されているかを、問題を提起するような視点から深く検討することが、構成の関連性を適切な度合いで実現するためのベストプラクティスとなります。特定の課題に取り組むために、生徒が他にどのようなスキルを必要とするかを考慮することは特に役立ちます。小論文やプロジェクトといったオープンエンドの課題では、その課題で評価の対象となる

リサーチスキルや分析スキルを実証するために、高度な執筆力が求められる可能性が特に高くなります。この点を内包しているのが、評価のユニバーサルデザインの概念です。これは、評価の対象と関係のない要素について困難を抱える生徒のために障壁を低減することを目指しています（Dolan et al., 2013）。

構成の関連性はカリキュラムと特に強い関係をもちます。多くの場合、評価における設問は、カリキュラムの教材のごく一部に基づいています。このカリキュラムの範囲が十分かどうかという問いは、構成の関連性の要素であり、個人の判断に基づきます。同様に、設問や課題の種類によってテストされる構成の種類も変わる可能性があるため、出題される問題の種類の選択も評価の設計の重要な要素です。

図 12
試験とカリキュラムの範囲



管理のしやすさ

管理のしやすさについては包括的な研究が実施されておらず、単一の定義や測定のアプローチが確立していません。大まかにいえば、「管理のしやすさ」とは、生徒、学校、IB のそれぞれにとって、評価を実施するうえで必要となる労力を指します。

生徒にとっての管理のしやすさは、多くの場合、評価を完了させるために必要な労力を指します。例えば、18 歳の生徒にとって、8 時間にわたる試験は過度な要求と考えられます。同様に、生徒にとっての管理のしやすさを評価する際には、特定の生徒にとって、特定の日のどの時間帯に評価が行われるか、また試験時間の長さはどのくらいか、という点も考慮する必要があります。IB は、IB 教育の一環として生徒が取り組むすべての IB 評価

の作業負担を総合的に考慮する必要があります。これは、個々の科目だけでなく、プログラムの妥当性を考慮するという原則に立ち返るものです。

学校が評価のための資材を提供しなければならない場合、評価の管理のしやすさに影響します。例えば、実務的な工学コースでは、各生徒に組み立て用のエンジンを提供する必要があるかもしれません。生徒の数によっては、管理できないほどのエンジンが必要となる場合もあります。

学校にとっての管理のしやすさのもう 1 つの側面として、生徒の成果物を採点のために IB に提出する方法が挙げられます。発表を録音、録画するという要件は、書面の成果物を提出する場合に比べて、学校に多くの負荷がかかります。

最後に、IB も、評価のために提出される成果物の量という点で、管理のしやすさを検討します。3 時間にわたる演劇作品を実際に鑑賞することは、生徒の能力を実証するエビデンスを得るうえでは最適な方法ですが、外部評価課題としては現実的とはいえません。また、幅広い選択問題を出題した場合、それぞれの問題の間で共通の基準を確立するために慎重かつ時間のかかる作業が必要になるため、管理のしやすさに影響する可能性があります。

管理のしやすさ、信頼性、構成の関連性の間には、しばしば対立関係が生じます。例えば、評価の作業量を増やす、または試験の時間を長くすることで、カリキュラム全体に対する生徒の理解について、より多くのエビデンスを得ることができます。これにより、甘すぎる採点や厳しすぎる採点が互いに打ち消し合う可能性が高くなるため、評価の信頼性が向上します。ただし、この場合、評価の目標そのものよりも、長時間の試験に集中して取り組み続けるという生徒の能力を試すことになるため、構成の関連性と、生徒にとっての管理のしやすさが損なわれます。

IB は、特に各コースの評価の合計作業量という点で、評価の管理のしやすさを厳密に制御しています。

公平性とバイアス

ある生徒にとって偶発的に有利に働く、または不利に働く場合、そのテストにはバイアスがあることになります。バイアスのある試験の例として、以下が挙げられます。

- すべてラテン語で書かれた「歴史」の試験
- すべての問いが、クリケットの試合で得点をとることに関連している「数学」の試験
- 床から 2 メートルの高さにイーゼルが置かれた「芸術」の試験

いずれの課題も、一部の生徒（例えば、最後の例では身長の低い生徒）にとってかなり不利に働く可能性があります。実際には、これほど明らかなバイアスはまれですが、注意を怠った場合、生徒の成績がバイアスによって大きく左右される可能性があります。バイアスを生じさせずに文脈に沿った問題を出題することには、特段の難しさがあります。IB の国際性を鑑みると、ある生徒にとってはなじみのある状況も、他の生徒にとっては非常になじみのない状況となる可能性があります。

「バイアス」は、測定される適性や到達度の差異とは関係のない、評価プロセスの結果における差異として定義できます。バイアスは、評価の**実施方法**、評価の**採点**（採点の信頼性の問題）、または評価**課題**そのものから生じる可能性があります。

評価の実施方法から生じるバイアス

このセクションの冒頭に示した例には、実施方法から生じるバイアスが含まれています。この場合、一部の生徒にとってイーゼルの位置が高すぎるという、実際には考えられないような状況によるバイアスです。評価の実施方法、特に試験の実施方法がバイアスを引き起こす例は数多くあります。特に一般的なのが、試験の実施時期に関連するバイアスです。世界の特定の地域において、気温が非常に高いまたは低い時期や大気汚染が深刻化する時期、または、生徒が試験に集中できない時期や十分な準備ができない時期に試験を実施することは、バイアスにつながります。これは IB にとって特に難しい課題です。世界中の国や地域（統治領）のニーズが拮抗するため、どの日付を選んだ場合でも、少なくとも一部の学校にとって不適切となる可能性があるからです。

他によくあるバイアスの原因として、試験会場の配置が挙げられます。例えば、直射日光の当たる席に座る生徒と、暗い席に座る生徒を思い浮かべてみてください。ある生徒は時間厳守を要求される一方で、別の生徒にはある程度の柔軟性が与えられるなど、試験規則の実践が一貫していないこともバイアスとなる可能性があります。IB は、各プログラムの『評価の手順』において明確で一貫した規則を定め、規則に違反した場合は不正行為として扱うことで、このバイアスを管理しています。

採点から生じるバイアス

採点から生じるバイアスは、そのほとんどが意図に反して発生する偶発的なものです。人間の意識は、意思決定を円滑にするために近道を選ぶ傾向があり、多くの場合、それが無意識のバイアスにつながります。IB には、このバイアスを軽減する義務があるものの、それに伴う罰則等は設けていません。

採点から生じるバイアスは、生徒の字のきれいさなどに対する個人的な印象（例えば、Hughes et al., 1983）、生徒の性別に基づく優遇措置（試験官が性別を知っている、またはある程度の確信をもっている場合）、書式、句読点、綴りなど、評価の状況によってはそれほど重要ではない要素に過度の注意が払われることなど、さまざまな理由で発生する可能性があります。この対策として、試験官の研修および採点作業の確認を実施することができます。

また、性別、国籍、学校に基づく無意識のバイアスも多くの文献で確認できます。これを最小限に抑えるため、IB は評価用に提出されたすべての課題を採点前に匿名化するよう努めています。

もう 1 つ、多くの研究によって示唆されているバイアスとして、「ハロー効果」が挙げられます。これは、生徒の最初の解答の質が高い場合、試験官は生徒について肯定的な印象をもち、その結果、後に続く問いにおいて生徒にとって過度に有利な判断を下すというものです。

IB では、試験官が特定のトピックに関する解答を公平かつ偏見なく採点できないと自認した場合に、提出された成果物の学術的価値に基づいて、バイアスのない状態で成果物を採点できる人物が成果物を評価できるようなプロセスを用意しています。

評価課題に関するバイアス

測定対象とは関係のない理由によって試験結果に違いが生じている場合、バイアスが発生しているといえます。例えば、数学の問題の背景として特定の文化の例が使われるケースにおいて、その文化的伝統（評価の要件ではない）に関する知識が不足していることが、理解やパフォーマンス（評価の要件）を阻害する場合などがこれにあたります。

心理測定情報に基づいたテストを作成する場合、事前テストで異常な解答特性が示された項目、またはさまざまな生徒のサブグループの間で大幅に異なる解答特性がみられた項目（「特異項目機能」と呼ばれる）は、バイアスがかかっているとみなされ、テストから除外されることがあります。生徒のサブグループは、性別、人種、社会階級、言語能力で定義したり、テストの対象である構成とは無関係だといえる特性によって定義したりすることができます。

特定の評価または評価要素にバイアスがかかっているどうかを判断する際は、評価の基礎となる構成にその課題がどのように明示的に関連づけられるか、また、バイアスをもたらし得る要因は何かを慎重に考慮する必要があります。純粋に統計的な根拠に基づいて判断を下してしまうと、バイアスのかかった評価と、生徒のサブグループ間の差異を明確にすることを目的とする評価を混同してしまう恐れがあります。ゴールドステイン（Goldstein, 1996）とハンフリーズ（Humphreys, 1986）は、客観的に決定された事実である「差異」と、差異の関連性についての判断である「バイアス」を区別することが有用であると示唆しています。

ブラック（Black, 1999）は、生徒に対する影響という意味で、設問が不公平となり得る 6 つの典型的な例を示しています。

1. 設問が設定された文脈は、その文脈になじみのある生徒に有利に働きます。例えば、米国に関連する文化への言及は、その他の地域の生徒と比較して、米国に暮らす生徒に有利になります。
2. 人間関係に関する小論文の問題は、豊かな感情表現が奨励される文化的文脈または家族的文脈をもつ生徒に有利になる可能性があります。
3. 多肢選択問題は男子が有利となる一方、
4. コースワークやプロジェクト課題は女子が有利になるかもしれません。
5. 特定の社会経済的背景に関連づけられた言語や慣習を使用する設問は、それとは異なる背景を持つ生徒に不利に働く可能性があります。
6. 特定の文化的背景の中でのみ理解される設問というものも存在します。例えば、独居の高齢者に関する問題は、一部の文化的文脈では非常になじみがないものの、別の文化ではなじみがあると感じられるかもしれません。

さらに、特に米国では性別によるバイアスのエビデンスが存在するものの、問題形式のどの側面がこの調査結果に寄与しているかは明らかになっていません。

評価の設計において、特定の評価を受ける生徒間の差異にどのように対応すべきかは必ずしも明確ではありません。評価は、さまざまな課題と問題の種類によって、バイアスの全体的な影響が軽減されるように設計する必要があります。明示的であるかどうかにかかわらず、文化や性別に基づく固定観念はいかなる形式であっても避けるべきです。不公平をもたらすことが知られている明らかなカテゴリーを避けるため、個々の問題の内容を精査する必要があります。また、さまざまな生徒の集団に対して試験問題の事前テストを行うことで、隠れたケースを特定できるかもしれません。ただし、バイアスを含む設問の種類や問題となる可能性のあるシナリオをすべて排除してしまうと、評価の設計者や問題作成者に残された選択肢はほとんどなくなります。その結果として生じる制約は、評価の妥当性に悪影響を及ぼすことになります。明白かつ不要な落とし穴を避けることとは別に、さまざまな種類の評価課題と形式を使用したバランスのとれたアプローチを評価の設計に採用することが、最も合理的な解決策となるでしょう。

また、生徒群において、異なる定義をもつ集団をいくつ考慮すべきかという問題もあります。例えば、学習スタイルがさまざまな異なる生徒を考慮する必要はあるのでしょうか。ヒエロニマスとフーバー（Hieronymus & Hoover, 1986）が述べているように、関心や動機の違いがバイアス要因だと考えられる場合、すべての課題や評価方法には、ある程度のバイアスが含まれると言うことができます。例えば、語学試験で使用されるテクストの抜粋が、ある生徒にとっては興味深いものの、他の生徒にとってはそれほどでもないという状況が考えられます。バイアスの回避を含む評価の公平性は、特に評価ツールにおける明らかなバイアスが訴訟につながる可能性がある特定の国や地域（統治領）では、大きな問題です。ところが、バイアスの証明は、パフォーマンスの差異とは対照的に、評価が実施される特定の社会的文脈に強く結びつく細かい判断の問題であることがよくあります。

ギップスとマーフィー（Gipps & Murphy, 1994）は、『A Fair Test? Assessment, Achievement and Equity（公平なテストとは　評価、到達度、公平性）』の結びとして、「公平なテストなど存在しないし、存在し得ない。状況は複雑すぎるし、その概念はあまりに単純すぎるからだ」と述べています。ただしこれは、評価の設計者や問題作成者が、バイアスや不公平の影響を減らすために全力を尽くすべきではないという意味ではありません。ギップスとマーフィーはまた、評価の設計者は、解答統計に応じて個々のテスト項目を操作することで実現される結果の平等ではなく、機会の平等と評価へのアクセスの平等を目標として設定すべきだという見解を維持しています。例えば、男子生徒の相対的な成績を向上させるために英語の試験に多肢選択式問題を導入することが、どこまで正当と見なされるのか、と疑問を呈しています。なぜならこうした対応は、「妥当性」という用語について広く受け入れられている定義に照らすと、評価の妥当性を歪めることになるからです。

評価プロセスにおける公平性の欠如は、教育の不平等に寄与する要因の１つにすぎず、おそらくそれほど重大な要因ではないということが広く認識されています。教育における不平等の原因は他にも多くあり、生徒の成果に大きな影響を与えます。その例として以下が挙げられます。

- 学校内での指導の質の違い
- 学校ごとおよび地域ごとのリソースレベルの違い
- 個々の生徒に与えられる社会および家庭からのサポートのレベルの違い

これらはいずれも、個々の生徒の成功に対して、いかに公平な評価プロセスであっても補うことができないほどの影響を与える可能性があります。例えば、スミスとトムリンソン（Smith & Tomlinson, 1989）は、生徒がどの民族かということよりも、学校がどれほど効果的かということの方が、試験結果の差を左右する大きな要因となることを発見し、異なる集団間の成績の差を修正するために評価ツールを調整しようとする試みは、時に不適切となる可能性があることを示唆しました。

このような検討は、到達度ではなく適性をテストすることの正当性を示す根拠となりましたが、現在では、社会的背景や教育経験を切り離して純粋に適性、能力、潜在能力のみを評価することは不可能であるという理解に至っています。また、教育的な到達度を、社会的および文化的背景から切り離して捉えることも不可能です。教育的な成功の概念は、特定の社会における限定された１つの部分の基準に従って定義・測定されます。

評価に対するバイアスと障壁を取り除く

バイアスに関するもう１つの課題として、評価課題が特定の生徒に不当な影響を与える可能性があるということが挙げられます。これは、評価への障壁を取り除くために必要な措置を実施できるような評価条件を整えることによって乗り越えられます。これにより、すべての生徒が他の生徒と同等の条件で教育的な到達度を実証することができるようになります。

本資料の「全員にとっての公正さを優先する」のセクションで、この考え方についてさらに詳しく説明しています。

バイアスを認識する

バイアスは、否定的にも肯定的にもなり得るということを念頭に置くことが大切です。ある課題が特定の生徒集団にとってなじみ深いものであったり、取り組みやすいものであったりする場合、それはやはりバイアスといえます。公正な評価のねらいは、すべての生徒に平等な機会を提供することです。

本資料ではこれまで、評価の設計や採点プロセスでバイアスが生じる可能性について説明してきました。評価サイクルのこれらの部分におけるバイアスに関して積極的に考えることは非常に重要ですが、潜在的なバイアスについて考えるだけでは十分ではありません。むしろ、試験全体と比較して当該生徒が特定の問題においてどのような成績を収めたかを分析することで、生徒の成績におけるバイアスのエビデンスを探す必要があります。

また、バイアスに関する意思決定を下す際は、固定観念ではなくエビデンスを根拠とすることが重要です。

同等性

同等性は、妥当性の側面の中でも特に複雑なものです。多くの場合、評価の結果は、選考のために生徒同士を比較する目的で使われます。2 人の生徒が同時に同じ試験を受けた場合、成績 7 を獲得した生徒の方が、成績 4 を獲得した生徒よりも、試験当日の成績が良かったと合理的に判断することができますが、これよりも複雑な比較が行われることもよくあります。そのような比較の例を以下に示します。

- 2 人の生徒が「歴史」で成績 6 を取得したが、それぞれ異なる設問に解答した、または異なる選択問題を選んだ場合
- 2 人の生徒のうち、一方は 2025 年 5 月に「スペイン語：文学」で成績 5 を取得し、もう一方は 2023 年 11 月に同じ科目で成績 5 を取得した場合
- 2 人の生徒のうち、一方は「物理」で成績 4 を取得し、もう一方は「化学」で成績 4 を取得した場合
- 2 人の生徒のうち、一方は「数学」で成績 4 を取得し、もう一方は「地理」で成績 4 を取得した場合
- ある地域の生徒が「コンピューター科学」で成績 3 を、「中国語：文学」で成績 6 を取得し、別の地域の生徒が「日本語：文学」で成績 5、「生物」で成績 4 を取得した場合
- 15 歳の生徒 2 人のうち、一方は MYP 修了証を取得し、もう一方は別の認定機関の資格を取得した場合。
- 2 人の生徒のうち、一方は「インドネシア語 B」の標準レベル（SL：standard level）で成績 6 を取得し、もう一方は「インドネシア語 B」の上級レベル（HL：higher level）で成績 5 を取得した場合。

同等性は、2 つの評価結果が何らかの意味で同等であるとみなせるかどうかを問うものです。IB ではさまざまな内容をテストし、それらに同等の価値があるかを問うことになるため、特に複数科目の間で同等性を判断することは簡単ではありません。

ここでは、抽象的な評価の特性ではなく、特定の目的に対する妥当性という概念が特に重要となります。生徒の「音楽」の成績は、プロの音楽家になる能力が身についているかどうかを示すうえで、「美術」の成績よりも効果的な指標ですが、歴史を学ぶ能力を示すという意味では、どちらの結果も同等である可能性があります。

生徒にはそれぞれ強みと弱みがあり、得意な科目と苦手な科目が存在するため、同等性の問題はさらに複雑になります。IB では、特に複数の選択肢からコースを選べるようになっている場合、受験者群がコースによって異なります。そのため、同等性は IB にとって特に難しい問題です。

IB は、同等性に関して以下の 3 つの原則を掲げています。

- 1 つの科目または学習分野における成績取得のための成果物の基準は、同じ年の複数セッションの間、および異なる年の複数セッションの間で同等性をもつ。
- プログラムの資格（IB ディプロマや MYP 修了証など）を取得するためのさまざまな道筋が同等となるよう、科目間または学習分野間の成績に一貫した意味をもたせる。

- 高次の思考スキルを評価するというIBの目標を追求する一方で、他の資格プログラムで取得された成績が、IBで取得された成績と同等性をもつよう注意を払う。

同等性を測定する

2つの評価の同等性を測定する方法は多数あり、このテーマに関する学術論文も数多く執筆されています。コウら（Coe et al., 2008）は、科目間の同等性に関する文献レビューにおいて、異なる科目の試験の難易度を比較する方法を、統計的手法と判断的手法という2つの大まかなカテゴリーに分類しています。

統計的手法は、評価における生徒の成績を比較し、傾向を見つけることに重点を置いています。これは、2つの評価が同等だとするならば、その2つの評価を受けた生徒の無作為サンプルを十分な数だけ用意した場合、その結果はおおむね同じになるはずだという考えに基づいています。

一方、判断的方法では、科目の専門家が多数の評価を検討し、その相対的な難易度について慎重に意見を導き出します。この際、同じ条件で比較できているかを確認するために、さまざまな調査ツールや手法が使用されます。

どちらのアプローチも重大な概念上の欠陥があると認識されており、コウら（Coe et al., 2008）はそれぞれの手法に対して6つの大まかな問題点を指摘しています。

表4

評価を比較するための統計的手法および判断的手法に対する批判

統計的手法に対する批判	判断的手法に対する批判
・ 研究において、指導や動機など、難易度以外の要因が測定される場合がある。 ・ 多次元性 — 評価対象が共通の特性をもたない場合がある。 ・ 非代表性 — 統計が本質的に偏った生徒集団に基づいているか。 ・ サブグループの差異 — さまざまに異なる生徒集団がさまざまな難易度を経験した場合、相対的な同等性の結論に疑問が生じるか。 ・ 手法の不一致 — いずれか1つの方法が「正しい」といえるか。 ・ 平等を強制することの問題点 — 資格取得を目指す生徒にどのような影響があるか。	・ 条件の幅広さ — さまざまな科目に適用できるようにするには条件を非常に広くしなければならず、したがって精度が損なわれる。 ・ 難易度に応じた評価 — 試験官は、難しい問いに対する不十分な解答よりも、簡単な問いに対する優れた解答に高い評価を与える傾向がある。 ・ 課題の種類に応じた評価 — 試験官にとって、短答式の問題と小論文など、種類の異なる評価課題を適切に比較することが難しい。 ・ 「判断的」手法でさえも統計的な比較によって支えられている（典型的な生徒の成績がどの程度になるかという経験に基づく）。 ・ 解釈と文脈 — 生徒の能力や公平性の判断評価では、単一の定期試験と一連のモ

統計的手法に対する批判	判断的手法に対する批判
	ジュール式評価との構造的な違いを考慮する必要がある。 ・ 総合的な判断 ― ほとんどの評価では複数の規準を測定するが、規準の間で適切なバランスが確保されている必要がある。

どちらの手法の支持者も、代替的なアプローチに対して非常に批判的であることが多く、また、現在のアプローチはすべて根本的な欠陥があると主張する人もいます。この問題を検討するために、コウら（Coe et al., 2008）は、イングランドの GCSE および A レベルの資格を対象に、科目間の同等性を測定する 5 つの方法を適用しました。その結果、科目間の同等性に関する統計的尺度と判断的尺度の間にはかなり高いレベルの一致があり、その差異は、科目間の差異よりもはるかに小さいと結論づけています。また、年ごとの相対的な科目難易度は安定していることもわかっています。

IB は、試験官の判断、統計分析、教師のフィードバックという 3 つの要素を活用して、年度の間の同等性、および選択問題の間の同等性を確保しています（「成績の付与と集約」のセクションを参照）。また、統計的アプローチと専門家による判断的アプローチの両方を通じて、科目または分野レベルでの同等性も確認しています。

妥当性に対する IB のアプローチ

- IB は、信頼性を最大化することよりも、構成の関連性をもつ現実に即した評価を行う方が重要であると考えています。
- バランスのとれた全人的な教育に重点を置く IB では、プログラムレベルで妥当性を効果的に主張できることを優先しています。これは、個別のコースの妥当性やコース内の選択項目の妥当性よりも重視されています。

妥当性を確保することは、多くの重要かつ相反するニーズにバランスよく対応することが求められる、複雑で多面的な行為です。妥当性がもつ相反する要素をバランスよく実現する方法は、1 つだけではありません。どこでバランスをとるかは、最終的には評価を開発する組織の価値観に基づく判断になります。

IB は、**実社会を反映**したうえで**重要な内容**をテストすることを重要視しています。「重要な内容」という点は、「構成の関連性」という言葉で表されます。つまり、IB の評価では、簡単に採点できる内容だけでなく、その科目にとって本当の意味で重要な内容が問われます。「実社会を反映」という点は、現実に即した評価を指しています。つまり IB 評価には、人工的で不自然なものではなく、実社会において、生徒が関連性の高い活動に従事する方法を反映した有意義な課題が含まれていることを意味します。

この 2 つの目標は、妥当性の他の側面、特に評価の**信頼性**を犠牲にして達成されます。現実に即した有意義な課題では、一般的に採点においてかなりの主観性が求められます。

例えば、明確な答えが 1 つしかなく、客観的に採点できる多肢選択式の評価と比べて、試験官の間のばらつきが大きくなることが許容されるということです。さらに、このような課題は評価の**管理のしやすさ**にも影響を及ぼします。IB にとって望ましい評価は、他の評価と比べて作成、管理、採点が難しく、時間もかかります。また、生徒に求められる努力のレベルも上がります。例えば、試験で高い点数をとることだけに集中するのではなく、有意義な研究課題に取り組むことになるため、生徒の作業量が増加します。

IB は、他の組織が評価の妥当性のバランスをとる際に、別の要素を優先する可能性があることを認めていますが、外部で検証される IB 評価に関しては IB の立場が適切であり、支持され得るものだと確信しています。

IB は、成功への意欲をもち、探究心と知識に富んだ思いやりのある若者を育成することで、他のカリキュラムをこえる成果を実現することを目指しています。「IB の使命」に概説されているように、IB はその教育プログラムを通じて、生徒が多様な文化への理解と尊重を育み、より良い世界の構築に貢献することを望んでいます。さらに、それぞれの IB プログラムは、IB が重視し育成を目指す 10 の人物像を特定した「IB の学習者像」に沿って、生徒を成長させることに力を入れています。

評価においてこれは、IB の目的が個々の科目や学問分野レベルではなく、プログラムレベルで定義されている、という点で重要な意味をもちます。したがって、評価の妥当性に関する問いはプログラムレベルで考えなければならず、個々の成績だけでなく、プログラム全体の資格の授与を規定する規則も含める必要があります。

これは、各コースや個々の評価課題を考慮することが重要ではないという意味ではありません。妥当性の要素の中には、このような詳細なレベルでしか意味を成さないものもあります。ただし、IB が全体的な妥当性の議論を行う際には、生徒が履修した学習プログラム全体を考慮します。

本資料で概説されている評価の原則のほとんどは IB 教育全体に適用されるものですが、「セクション C：IB プログラム固有のプロセス」では、各プログラムを順番に検討し、その固有の特徴について説明しています。

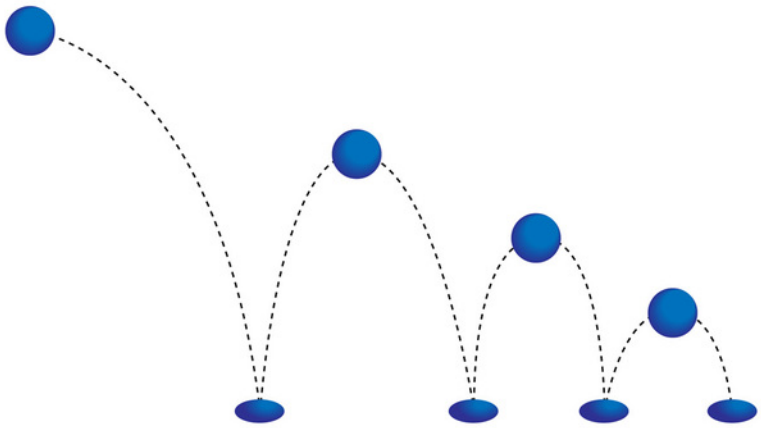
デジタル評価が妥当性にもたらすメリット

- 現代社会は紙ベースではありません。コンピューターデバイスが生活のあらゆる側面に浸透しています。
- 紙ベースの試験では、映像を提供することも、生徒との有意義なやりとりを実現することもできません。
- デバイスの汎用性により、試験の何か月も前に紙ベースの試験問題に対する特別措置を申請する必要がなくなり、設問に取り組むために必要な視覚的および音声的な調整をその場で行うことができます。
- いくつかの懸念はあるものの、デジタル評価は生徒の学問的不正行為と学校による不正または過失を、より効果的に防止することができます。

メール、テキストメッセージ、ソーシャルメディアは、コミュニケーション手段として手紙よりもはるかに一般的になり、コンピューターデバイスは多くの職場で日常的に使用されています。IB が現実に即した評価を目指すのであれば、その評価にテクノロジーとデバイスの使用を取り入れることは理にかなっています。

これまでの評価では、従来のテスト方式（紙と鉛筆の使用など）で評価できる内容には限界がありました。紙ベースの試験では、評価を操作したり対話したりする機会が与えられず、単純な刺激材料に対応したり問いに解答したりすることしかできませんでした。デジタル評価では、最も基本的なものであっても、刺激材料として音声や動画を含めることができ、試験監督者がクラス全体に対して教材を再生するのではなく、各生徒が個別にアクセスすることができます。MYP のデジタル評価では、物体がどのように動いているかを画像や長い説明で示すかわりに、アニメーション化された図やシミュレーションを適切な場所に含めることができます。

図 13
動画の使用による e アセスメントの設問の明確化



評価課題がより洗練されるにつれて、生徒が問題にどのように取り組むか、新しい情報にどのように反応するか、シミュレーションをどのように操作するかを、テクノロジーを使って評価できるようになります。AI を用いた個別指導システムでは、生徒の発話にデジタル評価が反応し、口述試験などの教師が実施する評価を再現できる可能性があります。この場合、生徒ごとに教師が異なるという問題は発生しません。デジタル評価の主な利点は、紙ベースの試験で評価できる内容に限定されるのではなく、IB が実際にテストしたい内容を評価するための新しい課題の設計が可能になることです。

妥当性のもう 1 つの重要な側面はバイアスを最小限に抑えることです。これは特に、評価を受ける生徒のために受験上の配慮が必要な場合にあてはまります。IB は、フォントの変更や用紙の色の変更に関する申請を定期的に処理しています。点字の試験問題の提供が申請されることもあります。

デジタル評価でこれらの障壁を完全に取り除くことはできませんが、フォントサイズや色をニーズに合わせて生徒自身が変更することは可能になります。また、スクリーンリーダーの機能をもつソフトウェアを使用する生徒にとっても、アクセシビリティーの面で大

きなメリットがあります。デバイスの使用によって、すべての特別な配慮に対応することはできないかもしれませんが、デジタル評価により、さまざまな生徒が追加のサポートを必要とせずに評価を受けられるようになります。また、必要な場合には追加のサポートも提供されます。

デジタル評価に関する最も大きな懸念は、テストのセキュリティーです。これは、妥当性を脅かすものでもあります。デジタル評価に対する批判として、紙ベースの試験よりも安全性が低いという主張が聞かれます。実際には、デジタル評価は紙ベースの試験が直面する一部のリスクからは守られているものの、他のリスクに対する脆弱性をはらんでいます。

デジタル試験の重要な利点として、試験問題を安全に送付できること、また試験問題へのアクセスを IB の規定時間内に制限できることが挙げられます。学校がアクセスを許可するまでユーザーは試験の内容にアクセスできません。これにより、紙ベースの試験を送付する場合と比べ、リスクを低減できます。紙ベースの試験は誰でも読むことができ、学校は試験の数日前から試験問題を安全に保管する必要があります。

デジタル評価では、生徒がインターネット接続機能をもつデバイスにアクセスできるため、試験の実施に関してさらなる課題が生じることは IB も認識しています。デバイスを介して認められていない資料にアクセスするリスクを軽減するため、IB のデジタル試験はロックダウンモードで実行され、試験中は他のコンテンツへのアクセスを禁止して、試験の安全性と学問的誠実性をサポートしています。さらに、生徒が設問にどのように答えたか（ログファイルなど）や、最終的な解答を記録する機能により、学問的不正行為に関連する行動を特定する事も可能です。

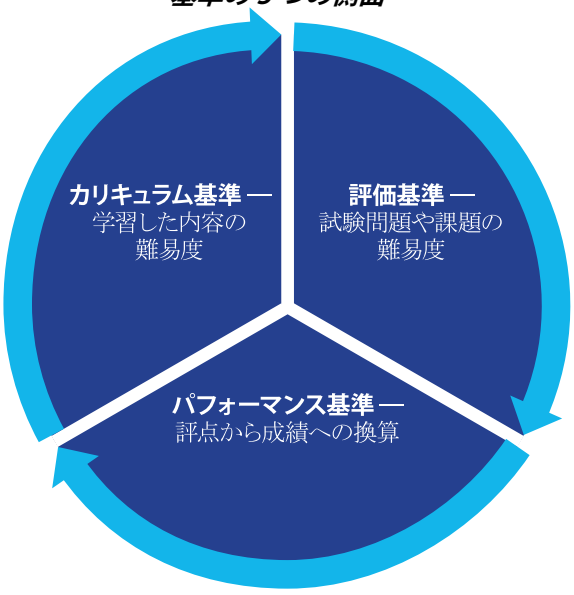
基準

- 基準の定義は、**カリキュラム**、**評価**、および**パフォーマンス**の 3 つの要素で構成されます。
- 形成的評価では、通常、カリキュラムと評価基準に焦点があてられます。
- IB は**到達度準拠**（以前は「軽度の目標基準準拠」と呼称）を使用します。これは、規準と過去の生徒の成果との比較とのバランスを使用して基準を設定することを意味します。
- 基準を維持することは、基準を設定することと同じくらい重要です。

基準の 3 つの側面

評価の文脈において、「基準」という言葉は通常、生徒のために設定された課題の難易度を表します。基準は同等性の核となる要素です。基準の概念は、3 つの方法で考えることができます。

図 14
基準の 3 つの側面



これらの各側面を変更することで科目の全体的な難易度を変更できますが、変更の影響やタイムスケールはそれぞれに異なります。

IB では、**カリキュラム**の変更は通常カリキュラムレビューにおいて行われますが、正式なレビューサイクル以外のタイミングで変更を行うことも可能です。公平性を保つには、教師がその変化に合わせて指導を調整する時間を与える必要があります。DP と CP では 2 年間（コースの期間）が理想とされる一方、MYP では、それより長い時間が与えられるこ

とがあります。これは、MYP が 5 年間に及ぶプログラムであるためです（ただし評価は通常最後の 2 年間に行われます）。

評価基準は試験開発サイクル（通常は 12〜18 か月）によって異なります。カリキュラム基準とは異なり、評価基準は年ごとに若干異なります。まったく同一の要求度をもつ試験問題を 2 つ設計することは不可能なためです。前回の試験と同じ問題を再度出題した場合でも、2 回目は生徒がその問いを知っていることになるため、難易度は低くなります。場合によっては、前回の試験が期待どおりに機能しなかったことなどを理由として、試験の評価基準を変更する決定が下されることがあります。必要となる変更の規模によっては、すでに開発中の試験問題が修正されることもあります。前述のとおり、公平性を保つためには教師に変更を通知する必要がありますが、変更の規模によっては、カリキュラムの変更よりも通知期間が大幅に短くなります。

パフォーマンス基準（すなわち、成績区分）は、評価基準の変更に合わせて毎年調整されます。例えば、試験の難易度が上がり、生徒の評点が全体的に低下した場合、同じレベルの生徒が前年度と今年度で同じ成績を取得できるように、成績区分が引き下げられます。他の 2 つの基準とは異なり、パフォーマンス基準を理解し、複数のセッションにわたって同じ基準を適用する必要があるのは、教師ではなく主に試験官です。成績区分が設定された時点で、パフォーマンス基準を変更することが可能になります。パフォーマンス基準が変更される場合、IB は学校に事前に通知し、学校が生徒の期待値を管理・調整できるようにします。このような場合、試験官が新たな基準がどのようになるかを理解し、それを将来のセッションに適用できるようにすることが非常に重要です。

評価の全体的な基準を設定するにあたっては、上記の 3 つの定義のバランスをとることが大切です。簡単な内容に対して非常に難しい問いを設定したり、簡単な設問群に対して非常に高いパフォーマンス基準を設定したりすることは可能ですが、その結果、構成の関連性が非常に低くなることがよくあります。例えば、テストで満点をとらなければ成績 7 が付与されない場合、生徒には非常に高いレベルの正確さが求められます。科目をよく理解しているものの、わずかに注意力に欠けている生徒や文章力が乏しい生徒が成績 7 を付与される可能性は低くなります。これは、評価の本来の目的にかなっているでしょうか。

基準の定義は、**形成的評価**と**総括的評価**の両方に適用されます。ただし形成的評価では、パフォーマンス基準の設定は通常、教師が生徒へのフィードバックを決定する際の判断の一環として行われます。形成的評価では通常、特定のテストが生徒にとって適切であること、また将来の指導に役立つ情報が提供されることを確認するために、カリキュラムと評価基準がより一層重視されます。

集団基準準拠と目標基準準拠

- 「集団基準準拠」とは、生徒のパフォーマンスの順位を出し、所定の割合に応じて生徒に各成績を付与する方法です。
- 「目標基準準拠」とは、生徒のパフォーマンス目標の説明に応じて、パフォーマンス基準を設定する方法です。

- IB は到達度準拠（「軽度の目標基準準拠」とも呼ばれる）を使用します。これは、規準のバランスと前年度の生徒の成果との比較を使用して基準を設定することを意味します。

集団基準準拠と目標基準準拠は、評価におけるパフォーマンス基準を設定・維持するための 2 つの異なる方法を表しています。

集団基準準拠

集団基準準拠はしばしば、標準テストと関連づけられます。原則としては、典型的な生徒群に対してテストを試行し、その結果（定義上、正規分布またはベル型曲線になる）を基準尺度として使用し、その後同じテストを受けるすべての生徒のスコアを生成します。最初の試行からスコアの標準分布を導き出すこのプロセスは、「規準化」と呼ばれます。

集団基準準拠を技術的に説明するとこのようになりますが、必ずしも固定分布がすべてのテストの結果に適用されるわけではありません。固定分布は、最初の規準化においてのみ使用されます。後からテストを受けた生徒のスコアの分布は、この正規分布と異なる場合があります。

実際には、集団基準準拠とは、生徒をパフォーマンスに応じて順位づけし、固定された割合に応じて各成績を付与するプロセスを指すことがよくあります。例えば、上位 15%の生徒に最高の成績が与えられます。

目標基準準拠

目標基準準拠型の評価は、グレイサー（Glaser, 1963）によって最初に提唱されました。これは、パフォーマンス基準の設定における大きな変化を表し、「明確に定義された行動領域に関して」生徒の到達度を測定することに重点が置かれています（Popham, 1978）。

目標基準準拠では、各学年で期待されることを表す説明文が事前に定義され、それを生徒のパフォーマンスと比較します。これは通常、科目または評価の専門家が専門的な判断に基づいて行います。

このアプローチの制約として、すべての専門家にとって明確であり、かつ意味が同一となる説明文を作成することが非常に難しいことが挙げられます。これについては、「たとえどんなに正確に表現された規準であっても、明確に解釈することはできない」という主張がなされています（Wiliam, 1993）。従来の目標基準準拠テストの結果は、関連領域の習熟度が示されたか示されなかったかの二択になります。

評価を実践するうえでは、どちらのアプローチにも重大な欠点があります。厳格な集団基準準拠では、現在のテストが最初のテストと同じ難易度であることを示す確固たるエビデンスが必要となる一方、目標基準準拠は、専門家の判断において問いの要求度が正確に考慮されないという、グッド&クレスウェル効果（Good & Cresswell, 1988）の影響を受けます。

到達度準拠

IB は「到達度準拠」(Newton, 2011) として知られるアプローチを使用します（以前は「軽度の目標基準準拠」(Baird et al., 2000) と呼称)。これは、規準に準拠する一方、グッド＆クレスウェル効果 (Good & Cresswell, 1988) のエビデンスも認識するものです。IB における実際の作業において、専門知識をもつ試験官は、規準（成績評価の説明）に基づいて、成績区分が設定される狭い範囲を確立するように求められます。そしてこれを前年度のパフォーマンスと一致するように計算された成績区分と比較します。成績はこの 2 つの区分の一致をもって設定されます。区分が異なる場合は、相反するエビデンスをどのように整合させるかについてさらに議論を重ねます。

最後に、目標基準準拠テストと集団規準準拠テストは、設問の種類ということよりも、生徒の解答の分析と解釈において大きく異なることを念頭に置くことが重要です。

基準の維持

- 適切な基準が設定されたら、それが毎年必ず適用されるようにします。
- カリキュラムの基準は、カリキュラムレビューにおいて再考されます。
- 評価基準とパフォーマンス基準は、専門的な判断と統計的エビデンスを組み合わせて維持されます。

同等性は、妥当性の重要な側面です。基準を維持するという観点から言えば、これは、毎年必ず同じ基準が適用されるようにしなければならないということを意味します。IB の基準は評点ではなく成績に基づいています。この 2 つの違いについては、本資料の「評点と成績の違い」のセクションで説明されています。

カリキュラムの基準はセッション間で変わらないため、最も簡単に維持することができます。評価はカリキュラム全体を正確に反映したものでなければならず、これは試験作成中に確認されます。ただし、外部要因によりカリキュラムの基準が変化する可能性があります。その典型的な例はコンピュータースキルです。デバイスが進化し、日常生活に浸透するようになると、かつては専門的で難しいと思われていた知識や理解が当たり前のものになります。このため、カリキュラムの内容が変わっていないにもかかわらず、このトピックに関連するカリキュラム基準は変更されています。

IB では、この問題に対処し、内容を最新の状態に保つために、カリキュラムレビューのサイクルを設けています。このレビューの結果、カリキュラム基準が変更されることが予想されます。その場合、生徒に不利益が及ばないように、評価基準とパフォーマンス基準のバランスをとる必要があります。

IB 試験の**評価基準**はセッションごとに変わります。2 つの試験が完全に同一の難易度になることはあり得ないためです。一部の教育システムでは、各設問の難易度を確立するために大規模な事前テストを実施し、生徒が経験する難易度について高いレベルの信頼性が確保されるよう、慎重に試験問題が構成されています。この方法は、受験前の生徒に問題が共有されるというリスクをはらんでいるため、IB では試験の事前テストは実施していま

せん。その代わり、経験豊富な試験作成者とチェッカーの専門的な判断を頼りに、一貫性がありバランスのとれた試験を作成しています。

IB はまた、全体的な基準が維持されるように、生徒の解答に基づいて**パフォーマンス基準**を調整します。パフォーマンス基準を維持するための 2 つの主な方法については、前のセクションで説明していますが、パフォーマンス基準を維持することこそが IB の成績授与プロセスの目的です。

総括的評価における生徒の成果の説明

- 成績は生徒の到達度を非常に簡略化した形で表すものですが、大学などの教育機関や雇用主といった関係者が、選考プロセスにおいて合理的な判断を下すことを可能にします。
- より複雑で総合的な情報のみが提供される場合、関係者は、意味のある選考を行うためにそのエビデンスを簡略化する必要がありますが、成績区分を設定する際に、IB ほど慎重な検討が行われない可能性があります。
- IB 評価では、生徒にとって有意義な結果を得るうえで、専門家の判断を使用することが重要な意味をもちます。IB は、公平性を保証するために、客観的かつバイアスのないエビデンスでこの判断を裏づける必要があることを認識しています。

成績の影響

優秀なシェフの知識、スキル、経験を思い浮かべてみてください。シェフがもつ知識とは、完璧にバランスのとれた料理を作るために必要な材料と味つけの組み合わせに関する知識です。シェフがもつスキルとは、食材の選択と下準備、調理、盛りつけとプレゼンテーションに関するスキルです。そしてシェフの経験とは、テクニックを駆使して完璧な一皿を仕上げること、食材が完璧に調理されているかどうかを判断すること、食材と料理の組み合わせのプレゼンテーションを通して最高の満足感を提供することです。

さて、こうした複雑な要素をすべて 7 段階という単一の成績に凝縮して、誰が最高のシェフかを決定するとしましょう。その結果にはほとんど意味がありません。たとえ尺度を 100 段階、あるいは 1000 段階に増やしたとしても、あまり役に立たないでしょう。差異を明確にできる余地は広がるかもしれませんが、異なるスキルセットを比較するという根本的な問題は、解決されないからです。

まさにこれが、評価において発生する問題です。生徒の知識、理解、能力の複雑さを 1 つの成績で最も効果的に表現するにはどうすればいいのでしょうか。たとえ生徒に関するすべての情報を正確に把握し、要約することが可能であったとしても、生徒の才能を完璧に反映した成績を付与することはできません。

この代替案としてしばしば提案されるのが、成績をまったく付与しないという考え方です。IB が各生徒に対して、出来がよかった部分と悪かった部分を説明する個別の文章を提供することはできるかもしれません。このようなアプローチは、優れた学習と教育の原則に沿ったものですが、生徒間の比較が非常に難しいという重大な欠点があります。

IB の評価の目的は、生徒の到達度を表現し、進学や就職をサポートすることです。これは、受け入れ機関（DP および CP の場合は大学が多い）が、どの生徒を受け入れるかを決定するために何らかの選考プロセスを実施する必要があるということを意味します。

> 他の規準の信頼性が試験よりも低い場合、それらへの依存を高めると、選考をめぐる意思決定の信頼性が低くなることは明らかです。
>
> (Cresswell, 1986, p. 42)

選考が実施される場合、IB は選考にかかる意思決定を可能な限り公正かつ有意義なものにすることで、IB の生徒を支える責任があります。IB が、受け入れ機関に対して生徒に関する説明的な情報のみを提供した場合、受け入れ機関は生徒を比較するための他の方法を見つける必要がありますが、その方法はほぼ確実に、試験結果を採点するという方法よりも信頼性や同等性が低くなります。

これは、総括的評価そのものが選考のための完璧な方法、あるいは特に優れた方法であることを示唆するものではありませんが、他の多くの方法と比べて公平性の高い方法といえます。さらに、他の関連指標（成績評価平均値、課外活動、個人エッセイ、推薦状など）と併せて使用することで、意思決定に役立つ情報を提供できます。そしておそらく最も重要なのは、評価を可能な限り公平で有意義、かつ信頼できるものにするために、IB が常に努力を重ねているという点です。

図 15
それぞれのペアの生徒たちの差異を明確にするべきか

図 15 の例では、最初の 2 人の生徒は一般的な DP の成績評価の説明における成績 5 と成績 6 をそれぞれ与えられています。必要な場合、この 2 人の生徒を比較して判断を下すことは公平であり、再度テストを受けても同じ結果になる可能性が高いでしょう。一方 2 つめのペアの場合、この違いは意味をなさない可能性が高く、2 人の生徒の能力はおそらく同等といえます。

この例から、すべての違いが意味をもつわけではないということがわかります。そして成績は、2 人の生徒の差異を明確にできると IB が考えている部分を示すものです。境界線上（すぐ上またはすぐ下）にいる生徒にとってはどちらの成績も公平となりますが、ほとんどの生徒にとっては、成績の違いは全体的なパフォーマンスにおける有意な差を表します。

このことから、成績を何段階に設定すべきかという問いが生じます。成績が 2 つ（合格または不合格）しかない場合、ほとんどの学生は公平な成績を得られますが、境界線上にいる生徒にとってその影響は非常に深刻です。対照的に、成績が 20 段階ある場合、はるかに多くの生徒が境界線上または境界線付近に位置することになりますが、一人ひとりの生徒にとって、正しくない成績を与えられることによる影響はそれほど重大ではありません。この概念は、クレスウェル（Cresswell, 1986）によってさらに詳しく考察されています。

IB では通常、評価によって提供され得る有意のカテゴリーの数、および境界線上の生徒の数と間違った成績を付与されることによる影響との間の適切なバランスを考慮して、**7 段階の成績**を採用しています。

専門的な判断の重要性

IB の評価の焦点となる複雑かつ高度な思考スキルは、判断に迷う余地のない単純な採点には適していません。生徒の解答は多岐にわたり、同じくらい有効かつ正しい解答が多数存在する可能性があります。研究によれば、複雑な知識やスキルは、小さな個別の構成要素に分解して指導すべきではないことが示唆されており、このような解答の採点にも同じ原則があてはまります。したがって、マークスキームを作成する際には、すべての生徒を一貫した方法で採点する方法とそれを実現する方法について、IB の科目の専門家がしっかりとしたガイドラインを提供する必要があります。

これは、試験官の専門的な判断、特に採点基準を設定し説明する上級試験官の専門知識を非常に重要視する必要があるということを意味します。これらを総合して考えた場合、結果の精度という文脈において、評価システムの信頼性に対する重大な（かつ克服しなければならない）課題が提起されます。

IB 評価の採点

- 採点とは、与えられた課題を生徒がどの程度達成したかを評価するプロセスです。
- 採点においては、解答の個々の詳細に焦点をあてることも、より総合的かつ全体的な判断を行うこともできます。
- 形成的評価の場合、採点結果は点数ではなく、記述文のみとなることもあります。

採点の定義

このセクションにおいて、「採点」および「採点方法」は、MYP、DP、CP の総括的評価で行われる IB 認定の採点を指します。形成的評価はすべてのプログラムで使用されていますが、多くの場合、これは教室で完了し、IB が試験官の採点を通じてその評価を認定することはありません。

採点は、生徒のこれまでの取り組みの優劣を説明するものではありません。生徒の到達度は、問題の難易度や各レベルの予想合格点など、さまざまな要素によって異なります。採点では、生徒の解答を期待される模範解答と比較します。

採点は、評価手法によって適切と判断される課題の性質と採点の種類に応じて、さまざまな方法で行うことができます。

採点方法

IB はさまざまな評価手法を使用しており、評価手法のニーズによって、適切な採点の種類を判断することができます。

場合によって、採点が非常に**客観的**、つまり生徒が正しいか間違っているかのいずれかのみとなることがあります。これは、解答に少数のキーワードしか求められない場合や、生徒が複数の選択肢から答えを選ぶ場合になどによく見られます。自動採点は、テクノロジーを使用して、事前に定義されたマークスキームに照らして生徒の課題を評価するプロセスです。ここでは、解答が客観的に正しいか間違っているかを判断します。IB では、解答が正解か不正解かの二択しかなく、解答が 100%正しいと確信するために試験官の判断を必要としない多肢選択問題に自動採点を採用しています。IB では、予想される解答の範囲は狭いものの人間の判断が必要とされる場合において、人間の代わりに人工知能による採点を行うことはありません。

その他のケースでは、採点ははるかに**主観的**なものとなります。生徒が許容可能な解答をしているかどうか、または「マークバンド」として知られる複数の記述の中で、どれが生徒の解答と完璧な解答の差異を最もよく表しているかの総合的な判断が、試験官に求められる場合があります。IB が使用するマークバンドの例が表 5 に示されています。

表 5

IB のマークバンドの例

評点	レベルの説明
0	成果物が、以下のレベルの説明に記されたいずれの基準にも達していない。
1〜3	・ 問題についての理解がほとんど示されていない。科目固有の用語が使われていない、またはその使い方が一貫して不適切である。 ・ 問題について最低限の説明しかなされていない。論点が表面的で、しばしば不明確である。 ・ 答案が説明的である。分析が表面的、またはまとまりがない。さまざまな視点についてまったく触れていない、またはわずかしか触れていない。結論が示されている場合、それが非常に表面的である、または答案の他の部分と整合していない。
4〜6	・ 問題についての基本的な理解が示されている。科目固有の用語は使われているが、不適切な使用が多い。 ・ 問題の説明が基本的で、十分に発展されていない。論点が不正確または不明瞭なことが多く、答案が伝えようとしていることがしばしば不明確である。 ・ 分析が限定的であり、答案全体が分析的ではなく説明的である。さまざまな視点について限定的にしか議論されていない。短絡的な結論が含まれている。
7〜9	・ 問題についてある程度の理解が示されている。科目固有の用語が、時おり適切に使用されている。 ・ 問題について十分な説明が成されているが、その説明は一部で明瞭性と発展性に欠けている。関連性の高い論点が述べられているが、正確さや詳細さに欠ける。 ・ 答案に分析が含まれるが、その分析は十分に発展されていない。さまざまな視点についてある程度議論されている。結論が含まれている。
10〜12	・ 問題についての深い理解が示されている。科目固有の用語が、おおむね適切に使用されている。 ・ 問題について説明は明瞭だが、さらなる発展が必要とされる。述べられた論点は関連性が高く正確だが、詳細さに欠ける。 ・ 答案に批判的な分析が含まれるが、その分析は十分に発展されていない。さまざまな視点について議論されている。答案が 1 つの結論に帰結し、その結論は提示された議論と一貫性がある。
13〜15	・ 問題についての非常に深い理解が示されている。科目固有の用語を適切かつ正確に使用している。

評点	レベルの説明
	・ 問題の説明は明瞭で、効果的に発展されている。論点は関連性が高く、正確で詳細である。 ・ 答案に、効果的に発展させた批判的分析が含まれている。さまざまな視点に関して、批判的な議論がなされている。答案が理路整然とした明確な結論に帰結し、その結論は提示された議論と一貫性がある。

もう 1 つ考慮すべき事項として、成果物のさまざまな側面（しばしば**規準**と呼ばれる）に対して、採点が個別に行われるかどうかという点があります。例えば、小論文は、文法の質、重要な事実の正確さ、小論文の構成、結論の質、という 4 つの規準に照らして測定されることがあります。このアプローチについては、規準を互いに独立させることが優れた実践となります。生徒が、解答の同じ 1 つの要素について複数の箇所で評点を与えられるという状況は公平ではないためです。

この**規準採点**の対極に位置するのが、**総合的印象評価**です。ここで、試験官は理想的な解答のさまざまな側面すべてを自らの中でバランスよく調整し、成果物を全体的に反映する最終的な判断を下します。

信頼性の高い採点を難しくする主な要因は、課題の性質と予想される生徒の解答の幅です。さまざまな解答が予想される幅の広い課題は、解答の範囲が限定される幅の狭い課題に比べて、採点の信頼性を確保することが困難になります。

さまざまな採点のアプローチについては、「採点」のセクションを参照してください。

形成的評価の採点

形成的評価は、すべての IB プログラムで使用されます。ただし PYP においては、これが**唯一の**評価形式となります。形成的評価では、採点に単純な数値を使う必要はありません。成果物を「完璧な」解答と比較し、類似点と相違点について説明的なコメントを提供することもできます。これは、「良い」とは何かを明確に示すために、生徒と共有する成功規準の形をとることもできます。この方法では、2 つの解答の良さを比較することは難しくなりますが、形成的評価のねらいは学習を支えるためのフィードバックを提供することです。

教室では、教師が形成的評価を使用して、生徒の知識、理解、スキルの足りない部分を特定することもできます。この場合、信頼性はそれほど重要ではありません。評点や成績をつける必要はなく、生徒が改善すべきトピックやスキルを列挙したものを評価の結果とすることができます。このような状況では、学習プログラムにとって重要な内容についてのフィードバックを生徒に与える必要があるため、構成の関連性が重要となります。別の言い方をすれば、生徒が知っていることを実証させるという点で構成の関連性が重要となります。

倫理的な考え方の育成

倫理的な考え方の定義

倫理的な考え方とは、**正直さ、敬意、誠実さ**の原則に従って、**責任をもって、うそ偽りなく**行動するという姿勢のことです。教育の分野では、正当かつ誠実な成果物を生み出すことの重要性を強調することで、学問的誠実性を育みます。この考え方は規則や方針を超越し、倫理的な意思決定と行動を学習コミュニティーの文化に根づかせます。そして、生徒が自らの学習プロセスを重視し、原則に基づいた選択を行い、公平な学問環境に貢献するよう後押しします。

倫理的な考え方を育む理由

倫理的な考え方を育むことは、教育の成果の信頼性と妥当性を確保するために不可欠です。また、評価の正当性を守り、生徒の成績が実際の能力を反映できるようにします。さらに、倫理的な行動を育むことで、生徒は卒業後の人生に備えることができ、高等教育や専門職の環境で成功するために不可欠な価値観を身につけることができます。倫理原則が無視されると、学校コミュニティー内の信頼が損なわれ、学校の信頼性が脅かされます。学校は、倫理的な考え方を育むことで、生徒たちが人生のあらゆる場面で公平さと尊重を重んじる、信念を持った世界市民になるようサポートします。

倫理的な考え方を育む方法

倫理的な考え方を育むには、コミュニティー全体にわたる総合的なアプローチが必要です。学校は、生徒の発達段階に合わせた期待事項を明確に伝えることを出発点として、学問的誠実性を学校文化に根づかせる必要があります。教師は、倫理的な行動の模範を示し、予防的なストラテジーを設計し、誠実性に関する議論をカリキュラムに組み込む、という非常に重要な役割を果たします。コーディネーターと学校リーダーは、方針が明確に定義され、一貫して適用され、すべての関係者に対する研修を通して支えられていることを確認する必要があります。保護者とのオープンなコミュニケーションは、期待事項を一致させ、家庭における倫理的な実践の重要性を強調する助けとなります。

定期的にワークショップを開催する、倫理的な課題の明確な例を示す、誠実性を実社会に応用することに関する対話を奨励するなどの取り組みを通して、これらの原則をさらに強化することができます。学校は、罰則のみに焦点をあてるのではなく、誠実であることの利点を強調することで、倫理的な意思決定に必要なスキルと理解を養おうとする生徒を支える環境を作り出します。

倫理的な考え方は単なる目標ではなく継続的な実践であり、責任の共有、一貫したメッセージ、学校コミュニティー全体の揺るぎない決意を通じて達成されます。より詳細な情報については、「倫理的な考え方の育成」（印刷可能資料）を参照してください。

評価における人工知能（AI）の倫理的な使用

人工知能（AI）が評価に及ぼす影響を議論する際には、以下の 2 つの点について考える必要があります。

- AI は評価対象にどのような影響を与えるか。
- 生徒について評価を下すプロセスにおいて、AI はどのような役割を果たすか。

AI は今や世界の一部となっています。そのため、AI と関わり、IB の生徒が確かな情報に基づいて倫理的に AI を使用できるように後押しすることが、責任ある行動といえます。

「これが私の成果物だと言うことは、何を意味するのか」

これは、AI が評価対象に与える影響という最初の課題を根底から支える重要な問いです。教育界が当初 AI に関して懸念していたのは、生徒が AI を使って「不正行為を行う」のではないか、つまり自分の代わりに AI に成果物を作成させるのではないかということでした。特定の評価において、AI が生徒よりも優れた成績を収めることができるという主張も数多くなされています。このような懸念は、「設問は何を目的としているのか」「その生徒についてどのようなエビデンスを集めようとしているのか」という、評価の最も基本的な問いに私たちを立ち返らせます。歴史的な出来事の日付などの事実はインターネットを使用すれば数秒で調べることができますが、時間的制約が厳しい状況、例えば飛行機を着陸させる場合などは、インターネットで答えを探すことが必ずしも適切であるとは限りません。

AI の他の用途についても同じことが言えます。発表形式ではなく、内容が重要となることもあります。この場合、コミュニケーションスタイルに対する評点は必須ではなく、生徒が自分にとって読みやすい形式で考えを提示できるよう、AI によるサポートを利用できます。一方、詩を書くときには、一つひとつの言葉選びが重要です。生徒が詩を完成させる際は、AI を一切使用すべきではありません。ただし、例えば類語辞典として使うことはできます。

AI の普及によって生じる重要な教育上の問いとして、「日常生活において責任をもって倫理的に AI を使用できるようにするために、生徒は何を学ぶ必要があるか」というものが挙げられます。この重要な問いは、時間の経過とともに進化していきます。現時点でのこの問いの答えとして、以下の点で生徒を教育することが挙げられます。

- 受け取った情報を批判的に捉え、バイアスが含まれている可能性や、結論が誤っている可能性を認識する。
- 情報源や参考文献をチェックし、正確に使用されていることを確認する。
- 自分が生み出したものとそうでないものを理解し、正直に提示する。

「こんにちは、私はこの試験の AI 評価アシスタントです」

AI は、生徒が評価される方法を少しずつ変えていきます。デジタル評価への移行により、従来よりもはるかにインタラクティブな課題が実施できるようになり、AI を組み込むことでこれがさらに発展します。例えば生徒は、デジタルアバターにアイデアを説明したり、対話したりできるようになります。このテクノロジーによって、現在では人間の試験官を

必要とするために管理が不可能な面接をより有効に活用できるようになるなど、評価に対して異なるアプローチをとる機会が生まれる可能性があります。

例えば、多肢選択問題やドラッグ＆ドロップ問題などの客観的な解答の自動採点は、長い間評価に使用されてきたツールですが、機械学習と AI の発達により、より主観的な問題や複雑な解答にも自動採点を使用できる可能性が広がっています。当面の間は、人間の試験官が重要な最終評点を決定する必要がありますが、AI は採点の質をサポートし、もう 1 人の試験官によるダブルチェックが必要とされる成果物を特定できるという点で、大きな可能性を秘めています。

AI による採点は、授業の一環として実施される形成的評価をサポートすることもでき、生徒の過去の解答に基づいて、生徒の能力に合わせて問題を調整する適応型評価の機会につながる可能性があります。このアプローチはこれまで多肢選択問題に使用されていましたが、AI の進歩によって、さらに幅広い種類の問題に適用され、生徒に合わせて個別最適化した試験を提供することが可能になります。

新しいテクノロジーの影響を考慮する場合、テクノロジーを導入してもなお、評価の目的が意図された学習成果に関連しているか、そして評価そのものがその目的に対して妥当であるかを確認することが非常に重要な要素となります。

繰り返しとなりますが、生徒が**学習サポート**の目的で AI を使用することは有益である一方、**学習の代わりに** AI を使用したり、自らが執筆したものではない学習成果物を提出したりすることは認められません。

関連文献

- 「人工知能（AI）ツールの使用について説明する際の重要なポイント」（1 ページのファクトシート）
- IB 資料『生徒のコースワークにおける人工知能（AI）を評価するための 13 のシナリオ（通常版）

実践の定義

- このセクションで説明されている実践は、MYP、DP、CP における総括的評価に適用されます。
- IB が総括的評価を実施しない PYP や、学校が実施する形成的評価には適用されません。

このセクションでは、外部評価やモデレーションを受ける生徒に対して、IB が評価結果を作成するための実践方法について概説します。PYP における教師作成の評価など、IB が実施しない評価は対象外とします。

原則とは、IB の行動や実践の**理由**を表し、**実践**とは、IB が物事を行う**方法**を説明するものです。このセクションでは、IB が評価結果の妥当性を確認するために行う高次の実践について説明します。

その下に位置するのが IB の**手順**です。これは、それぞれの実践にかかる一つひとつのステップを説明するものです。これらのプロセスが学校に関係する場合、各プログラムの『評価の手順』に詳細が記載されています。

IB が測定する対象および事前学習の役割

総括的評価の目的は、生徒が MYP、DP、CP の学習を完了するにあたり、**知識**、**理解**、**スキル**の観点からそのパフォーマンスを測定することです。これは、例えば以下のような意味をもちます。

- 総括的評価の結果は、その時点における生徒のパフォーマンスを反映したものです。総括的評価は、生徒の潜在能力や学習の進捗状況を測るものではありません。
- IB の評価は、生徒が特定の設問や特定の日に限って、期待されたパフォーマンスを発揮できなかったことによる評価の不正確さを最小限に抑えるように設計されています。これは通常、生徒が自分の能力を示す機会を複数回得られるよう、複数の試験を実施することにより達成されます。ただし試験の回数は、管理のしやすさを損なわず、かつ生徒に過度の負担がかからないように設定する必要があります。
- 総括的評価は、生徒の集団ではなく、個々の生徒の知識、理解、スキルを表します。
- IB が評価の成績を付与する際、事前の学習経験は考慮に入れません。つまり、生徒がその IB プログラムを開始する前に取得した資格、成績、実績は考慮しないということです。科目または学習領域と事前学習との関連性については、IB の各科目の『指導の手引き』で説明されています。

生徒の到達度を報告する

- 「IB の使命」の焦点は、より良い世界を創造できる若者の育成を可能にすることです。効果的な評価とは、この目標を支えるものであるべきです。
- IB の成績には意味があり、成績区分はこの意味を考慮して設定されます。
- 成績は生徒の到達度を非常に簡略化した形で表すものですが、大学などの教育機関や雇用主といった関係者が、選抜に関して合理的な判断を下すことを可能にします。
- より複雑で総合的な情報のみが提供される場合、そのエビデンスを簡略化して意味のある選考を行う責任は IB 以外の人や組織に課せられ、成績区分を設定する際に、IB ほど慎重な検討が行われない可能性があります。
- 生徒の到達度は試験の結果だけで測ることはできません。また、たとえ成績を扱う場合であっても、ある生徒にとっては残念な結果が別の生徒にとっては素晴らしい成果となることがあります。

国際バカロレア（IB）は、多様な文化への理解と尊重の精神を通じて、より良い、より平和な世界を築くことに貢献する、探究心、知識、思いやりに富んだ若者の育成を目的としています。

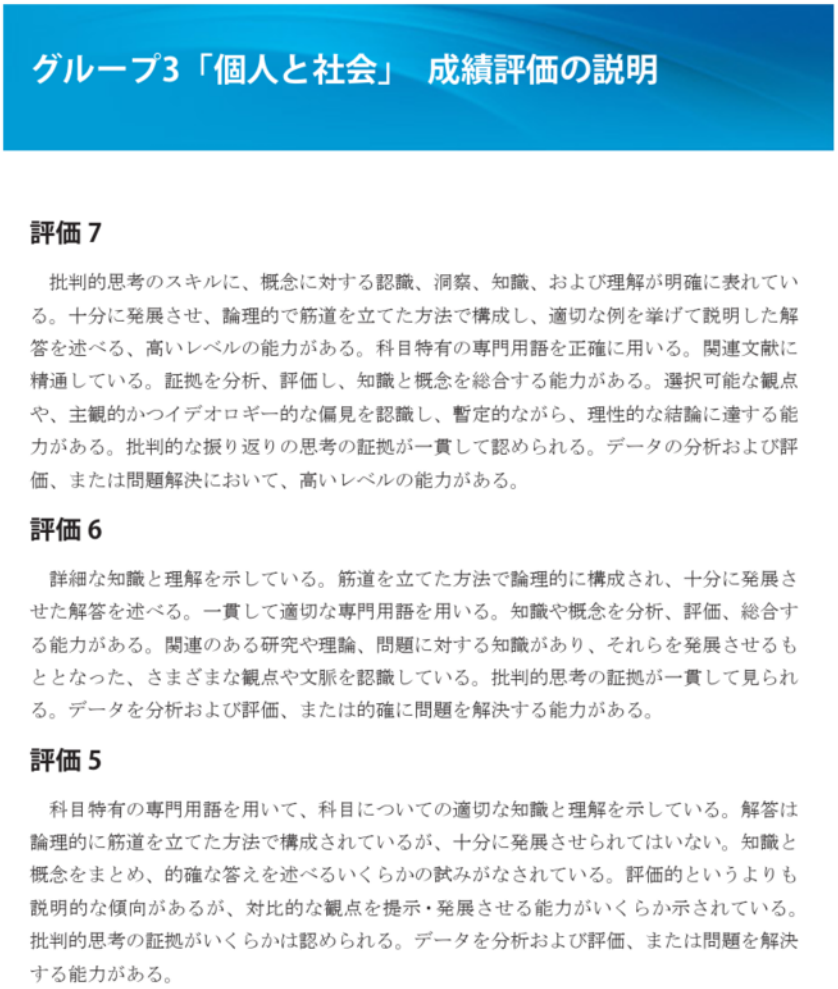
（「IB の使命」、2024）

IB の成績の意味

IB の評価の結果は、成績として表されます。この成績は、解答における生徒の取り組みの水準を示すものです。

IB は、各成績について説明を公開しています。IB が生徒に求める水準は年齢によって変わるため、この成績評価の説明は MYP、DP、CP でそれぞれ異なります。

図 16
DP の成績評価の説明の例



グループ3「個人と社会」 成績評価の説明

評価 7

批判的思考のスキルに、概念に対する認識、洞察、知識、および理解が明確に表れている。十分に発展させ、論理的で筋道を立てた方法で構成し、適切な例を挙げて説明した解答を述べる、高いレベルの能力がある。科目特有の専門用語を正確に用いる。関連文献に精通している。証拠を分析、評価し、知識と概念を総合する能力がある。選択可能な観点や、主観的かつイデオロギー的な偏見を認識し、暫定的ながら、理性的な結論に達する能力がある。批判的な振り返りの思考の証拠が一貫して認められる。データの分析および評価、または問題解決において、高いレベルの能力がある。

評価 6

詳細な知識と理解を示している。筋道を立てた方法で論理的に構成され、十分に発展させた解答を述べる。一貫して適切な専門用語を用いる。知識や概念を分析、評価、総合する能力がある。関連のある研究や理論、問題に対する知識があり、それらを発展させるもととなった、さまざまな観点や文脈を認識している。批判的思考の証拠が一貫して見られる。データを分析および評価、または的確に問題を解決する能力がある。

評価 5

科目特有の専門用語を用いて、科目についての適切な知識と理解を示している。解答は論理的に筋道を立てた方法で構成されているが、十分に発展させられてはいない。知識と概念をまとめ、的確な答えを述べるいくらかの試みがなされている。評価的というよりも説明的な傾向があるが、対比的な観点を提示・発展させる能力がいくらか示されている。批判的思考の証拠がいくらかは認められる。データを分析および評価、または問題を解決する能力がある。

この一般的な成績評価の説明は、教科内のすべての科目で同じである必要がありますが、IB では、それぞれの状況で何を意味するのかを理解しやすくするため、この説明を科目固有の文脈にあてはめることが多くあります。この科目ごとの文脈によって評価の水準が変わることはない、ということを覚えておくことは重要です。「言語」の成績 4 は、「理科」や「芸術」科目の成績 4 と同じ意味をもつ必要があります。これは、すべての成績が同等の重みをもつという IB のプログラムのアプローチにおいては、ごく自然なことです。ただし、このような概念が意味を成すかどうかについては、教育の専門家の間で多くの議論が交わされています ― 2 つの科目の到達度をどのように比較できるのか、そもそも、そのような比較を試みることに意味はあるのか、と。この概念については、「同等性」のセクションでさらに詳しく探究しています。

図 17
この 2 つの成果物をどのように比較できるか。

9. (a) $x = e^{3y+1}$

両側の自然対数をとると

$$\left(f^{-1}(x) \right) = \frac{1}{3} (\ln x - 1)$$

(b) Qの座標は(1,0)

$$\frac{dy}{dx} = \frac{1}{x}$$

Qでは, $\frac{dy}{dx}$

$y=x-1$

(c) 求められる面積をAとすると

$A=\int_1^e 1dx-\int_1^e \ln x dx$

部分積分を使って$\int \ln x dx$を求める

$$= \left[\frac{x^2}{2} - x \right]_1^e - \left[x\ln x - x \right]_1^e$$

$$= \frac{e^2}{2} - e - \frac{1}{2} \left(\frac{e^2 - 2e - 1}{2} \right)$$

異なる科目における IB の成績の同等性をめぐる議論は、妥当性に関する検討事項の一環であり、IB の成績の目的に依存しています。IB の成績は、関係者が生徒の到達度を比較できるようにすることを目的としています。したがって、統計的および定性的な方法を使用して成績の意味の平等性を目指すことには大きな意味があります。

評点と成績の違い

評点と成績は同じものではありません。

図 18
評点と成績を比較する例え

「優れた」受験者には課題の大部分を完了することが期待される

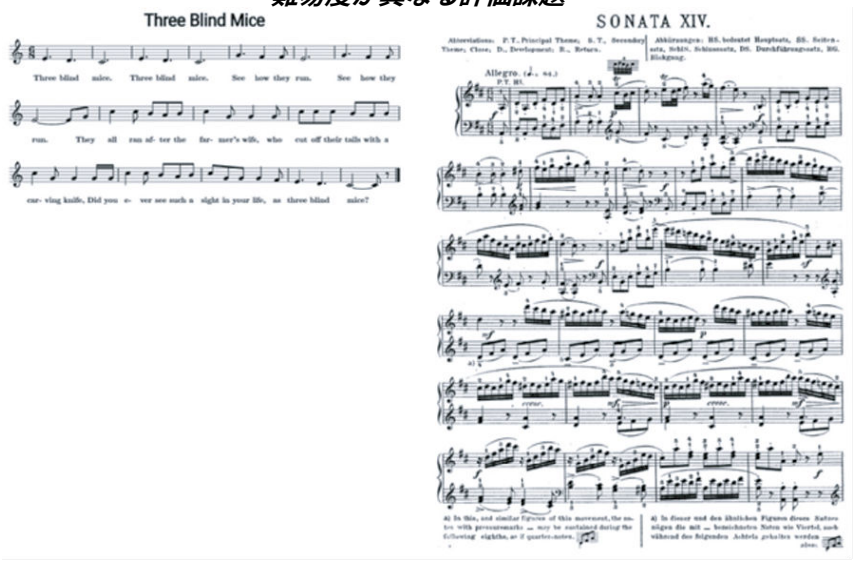
「優れた」受験者には課題の一部のみを完了することが期待される

評点と成績の違いを説明するために、多くの例えを使うことができます。例えば、図 18 では、歩いた距離を評点として表すことができます。これは、人が移動した距離を示す共通の尺度です。ただし、所定の距離を歩くことがどれほど大変なことなのかを理解するには、その人が歩いている場所を考える必要があります。成績を設定する際には、この点が考慮されます。

- **評点**は、生徒がある課題を**どの程度**完了させたかを表します。
- **成績**は、その生徒の評点が表す到達度を数量化するため、その課題の**難易度**を考慮に入れます。

図 19 に示されている 2 つの例を考えてみましょう。最初の例では、16 歳の生徒が「良い」成績を取得するには、ほぼすべての設問に正解することが求められます。2 番目の例では、生徒が「良い」成績を取得するために正解しなければならない設問数が、1 番目の例よりもはるかに少なくなります。

図 19
難易度が異なる評価課題



「優れた」受験者には課題の大部分を完了することが期待される

「優れた」受験者には課題の一部のみを完了することが期待される

この検討は、質の高い評価の設定をめぐる課題の 1 つにつながります。生徒には、もてる力をすべて発揮する機会を与える必要がありますが、課題が簡単すぎる場合、それが不可能となることがあります。反対に課題が難しすぎると、多くの生徒が課題にとりかかることすらできない状況が生じる可能性があります。この場合、成績を使用して生徒の差異を明確にすることが難しくなります。

あまりにも簡単な課題は、理解度ではなく正確さを測ることにつながる可能性があり、その結果、トピックをよく理解しているにもかかわらず、細かいところを間違ってしまったために最高の成績を得られなくなるという事態が生じることがあります。

評点と成績の違いについて詳しくは、「The difference between marks and grades: How we take into account the difficulty of an assessment, and not just how much a student got right（評点と成績の違い：生徒の正答率だけでなく評価の難易度も考慮する」（動画）をご覧ください。

効果的な試験セッション

図 20
効果的な試験セッションに対する見解の違い

すべての設問に対する答えがわかった

コースにふさわしい生徒を選ぶことができた

自校の受験者は近隣の学校よりも高い成績を収めた

設問や採点にミスがなかった

何をもって成功とするかは、見る人の視点によって大きく異なります。IB は、高次の構成の関連性と、試験の実施のしやすさの両方に焦点をあてています。効果的な試験セッションとは、以下の結果を生むものです。

- 評価によって、すべての生徒がその能力を示す公平な機会が与えられた。
- 学校および生徒にとって、複雑でわかりにくい部分がなかった。
- すべての関係者が、IB が公開した結果（成績）への信頼性を維持することができた。

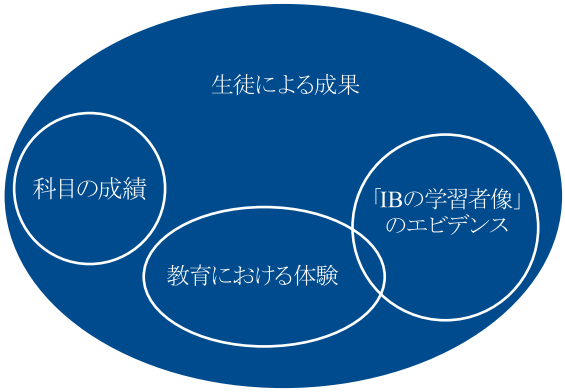
成績と到達度

「IB の使命」に示されるように、IB 教育の目標は、単なる学業成績をはるかにこえるものです。これは、「IB の学習者像」に反映され、IB 資料『国際バカロレア（IB）の教育とは』（2019 年発行）において明確に説明されています。

評価の結果は、この「IB の使命」のごく一部のみに焦点をあてています。仮に、「思いやり」や「挑戦すること」といった生徒の重要な資質に数値を与えることが妥当であるとした場合、IB の試験で使用される補完モデルを踏まえて、数学の評価においてこれらの資質

にどの程度の配点を割りあてるべきでしょうか。評価によってこのような目標を支える肯定的な逆流効果をもたらすことも、IB の評価の原則の 1 つです。

図 21
生徒の到達度を包括的に示す図



IB プログラムの成果を報告する際には、生徒の到達度全体を反映するために、評価成績以外の要素も考慮に入れることが重要です。

IB は評価における生徒の最終的な到達度のみを記録し、その結果を達成することが生徒にとってどれほど困難であったかや、生徒がもつ潜在能力については一切示しません。この 2 つの要素の重要性は認識しているものの、総括的評価の中でこれらを有意義に測定することは不可能であると考えています。こうした評価は、生徒を総合的に理解できる立場にある学校が果たすべき責任です。

生徒の到達度の「付加価値」尺度や「過去の到達度に基づく予測スコア」を計算する方法はありますが、生徒の成功の指標としてそれらを使用することには慎重を期する必要があります。このような尺度は平均的な生徒を基準としていますが、各生徒がもつ個性と特性の唯一無二の組み合わせは、プログラム全体を通じてその生徒とともに学ぶ機会を得た人々によって、価値あるものとして適切に評価されるべきです。

予測スコア

IB 資料『プログラムの基準と実践要綱』(2020 年版）では、「学校は、学問的誠実性を重視し、可能なかぎりの正確さで算出した予測スコアを（中略）IB によって認証された評価の一貫として IB に伝えること」と記載されています。

生徒の成績は学習過程を通じて変化する可能性があるため、最終成績を予測するのは難しい場合があります。予測スコアにはいくつかの重要な機能があります。成績授与会議において、IB が各学校の受験者群の全体的な強みと各科目の成績分布を理解するためのエビデンスを提供します。また、予測される成績と実際の成績の間の大幅な差異を特定する手がかりとなり、必要に応じて追加の品質チェックを実施できるようになります。学校にとって、予測スコアは生徒とその保護者の期待レベルを設定し管理するうえで役立ちます。

教師が生徒全員の成績を完璧に予測する可能性は低いものの、できる限り正確に予測できるようにすることが目標です。予測スコアが正確であればあるほど、プロセスにおける有用性と建設性が高まります。

予測スコアは評価にとって重要なデータポイントであり、科目の成績区分を設定するために科目全体レベルで考慮されるほか、その正確さを示すエビデンスがある場合には、学校レベルでも考慮されます。したがって、予測スコアが正確であり、信頼できるデータポイントとして見なされることが重要です。

予測スコアには、以下のような特徴もあります。

- 学校のためのデータポイントであり、生徒と共有するかどうかは学校の裁量に委ねられる。
- プログラム修了後の進路に関する能力を示す指標として、生徒の成功を支える。
- 評点ではなく成績（1〜7、A〜E）で表される。

予測スコアを算出する際は、予測スコアにあてはまらないものを理解することが重要です。以下は予測スコアにあてはまりません。

- 特別な事情に対する調整を加えるための仕組み
- 処罰のためのツール
- 前回セッションの成績区分に基づく公式
- 大学入試用の予測スコアといった、他の用途と混同されたもの
- 生徒や保護者をなだめるための仕組み
- 能力を過大評価または過小評価したもの

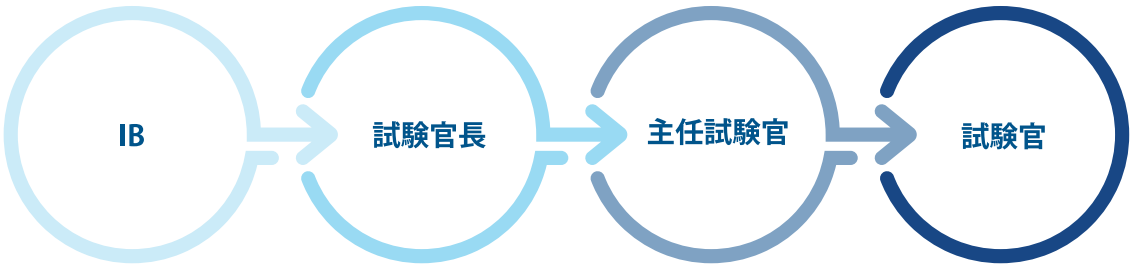
予測スコアの詳細と正確さと向上させるための方法については、印刷可能な資料である「予測スコア：教師用ガイド」（PDF）と、次に示すプログラム・リソース・センターの追加ガイダンスを参照してください。

- MYP resources（MYP リソース） > MYP eAssessment（MYP e アセスメント） > Cross-session resources（セッション共通リソース） > Predicting IB Grades（IB の成績の予測）
- DP resources（DP リソース） > Assessment（評価） > Cross-session resources（セッション共通リソース） > Predicting IB Grades（IB の成績の予測）
- CP resources（CP リソース） > Assessment（評価） > Cross-session resources（セッション共通リソース） > Predicting IB Grades（IB の成績の予測）

評価のプロセス：役割と責任

- 評価プロセスにおけるそれぞれの役割には、独自の責任とスキルセットがあります。
- 場合によっては、評価サイクルの異なる時点で同じ人物が異なる役割を果たすことがあります。
- これらの役割は、IB によって遂行されるものもあれば、IB コミュニティーの専門家が担うものもあります。後者の場合、最終的な承認と評価における当該要素の質の維持は、IB が責任をもって行います。

図 22
評価サイクルにおける重要な役職の責任と義務



IB
- 評価のあらゆる側面に対して責任を負う
- 試験官の採用、試験官の質、成績開示などの評価プロセスの実施を担当する
- 生徒の学問的不正行為、学校による不正または過失、受験上の配慮と特別な事情に関する問題について決断を下す
- 試験官長が推奨する成績区分を受諾する、または異議を唱える

試験官長
- 担当科目（教科）のすべての評価要素を監督する
- 試験の作成を含み、すべての評価要素が同じ基準に沿っていることを確認する
- 評価に関する学問的な問題の裁定を下す
- IBに最終的な成績区分を推奨する

主任試験官
- 1つの評価要素を担当する
- その評価要素において生徒の解答に付与される評点に関する最終的な決定権をもつ
- その評価要素においてすべての試験官が採点基準を理解していることを確認する
- その評価要素の成績区分の設定において試験官長に助言を与える

試験官
- 主任試験官が設定した基準に従って生徒の成果物を採点する

主任試験官と試験官長

主任試験官：

- 内部評価など特定の評価要素を管理する。
- 同一セッション内および複数のセッション間の採点基準を設定し、標準化の議論や品質モデルを通じて試験官に採点基準を説明する。
- 担当する評価要素における推奨成績区分について、試験官長に指針を与える。
- 利益相反がない限り、通常は評価コンテンツの開発者も兼任する。
- 各試験セッションにおいて外部教育専門家として IB の業務に従事し、通常は複数のセッションにわたってその職務を務める。

試験官長は、関連する複数の評価要素の質を維持する責任を負います。DP と CP では単一の科目全体を担当し、MYP（通常、分野ごとに 1 つの評価要素のみ）では、「理科」や「言語の習得」といった教科を担当します。試験官長は、この分野における IB の学問的専門家の役割を担います。

試験官長：

- 科目全体（DP および CP）または教科（MYP）に関連する評価要素の全体的な質を保証する。
- IB の学問的専門家としての役割を担い、基準を維持して一貫性を確保する。
- 対立の解決をサポートする。
- 成績付与のプロセスを主導し、成績区分を IB に推奨する。
- 特定の評価要素の主任試験官を兼任することも多い。
- 特定の期間にわたって IB の業務に従事する。

履修者が少ない科目の場合、IB は試験官を首席試験官の役割に任命します。履修者が少ない科目は主任試験官や試験官の数が少なく、首席試験官は、自分の担当科目に関して試験官長と同じ責務を担います。試験官長はまた、カリキュラムレビュー活動を通じて、カリキュラムと評価に関する議論と改善に向けて IB に協力するよう求められます。カリキュラムレビューについては、本資料では説明されていません。

その他の試験官の役割

試験官は、評価のために提出された課題を主任試験官が設定した基準に従って採点する責任があります。この基準を理解していること、そして品質モデルを通じて基準を適用していることを証明する必要があります。

セッションで採点を担当したい試験官は、IB に応募します。採点を希望する科目の専門家でなければならず、通常はその科目の教師が務めます。試験官は、対象の年齢層の生徒を指導した経験を有していなければなりません。IB は、推薦者の提供を含む厳格な応募プロセスを通して、その資格を検証します。通常、試験官には、採点対象として 1 セッションあたり 1 つのコースワーク要素と 1 つの試験要素のみが割り当てられます。この 2 つの要素の採点期間は重複しません。試験官は生徒の「ライブ」スクリプト（実採点対象の答案）1 件ごとに報酬を受け取ります。ライブスクリプトには、認定用スクリプトやシードスクリプト（「品質モデル」のセクションを参照）は含まれません。認定用スクリプトや

シードスクリプトは、試験官が採点基準を理解していることを IB に示す目的で用意されたものです。

特に経験豊富な試験官はチームリーダーを務め、IB からの依頼に基づき、他の試験官が正しい採点基準を理解できるようサポートします。チームリーダーは、資格認定プロセス全体を通じて試験官をサポートし、シードスクリプトの採点が確定基準から大幅に逸脱している場合にはフィードバックを提供します。科目ごとのチームリーダーの数は、試験官の全体数、ひいては受験者数によって決まります。履修者が少ない科目の場合、主任試験官がすべての試験官にこのサポートを提供する場合があります。

標準化チームのメンバーは通常、品質モデルの開発につながる定性的な議論に関与する、経験豊富な試験官から構成されます。標準化チームのメンバーは主任試験官の基準の適用について議論し、その基準が確実に適用され、遵守されるようにします。

試験官の役割のさらなる詳細については、「Introducing: The examiners: Who are they and what do they do? (試験官の紹介：試験官とは誰か、何をするのか)」（動画）を参照してください。

IB スタッフの責任

サブジェクトマネージャーと評価運営アナリストの仕事は、IB の観点から信頼性の高い評価を確実に実施することです。

IB 内の各チームが、試験問題の調整などといった受験上の配慮の申請、特別な事情に関する申請、学問的誠実性に関する問題、および試験官の質と学校のモデレーション係数のモニタリングを担当します。

評価担当ディレクターは、評価開発・実施の担当責任者、および評価の原則と実践要綱の担当責任者のサポートを受けながら、各科目の成績付与会議の結果として試験官長によって提言された成績区分を検討し、その成績区分の承認、または、試験官長への再検討依頼を行う責任を負います。

最後に、評価サイクルの一環として下されたすべての決定の責任は、IB が負います。外部の専門家を使うことによって公正かつ質の高い評価の実施がサポートされますが、最終的な説明責任は IB にあります。

試験の作成における役割

評価コンテンツ開発者は、試験作成プロセスにおいて指定された役割を担います。この役割には、作成者、校閲者、チェッカーの仕事が含まれます。評価コンテンツ開発者に必要な主なスキルは以下のとおりです。

- 創造性：カリキュラムに整合する魅力的な試験問題を作成する。
- 文化的認識：試験問題にバイアスがかからないようにする。
- コミュニケーションスキル：明確であいまいさがなく、翻訳しやすく、意図されるスキルをテストでき、信頼性の高い採点につながるような問題を作成する。

評価の作成に関係するその他の役割の例として、以下が挙げられます。

- テクニカルデザインエディター：試験問題の体裁を整える

- 翻訳者：求められる言語に評価を翻訳する
- 外部校閲者：生徒の立場から評価に取り組むことで、長さや明瞭さに関するフィードバックを提供する

評価の作成に関するさらなる詳細は、「評価の作成」のセクションを参照してください。

試験官の序列

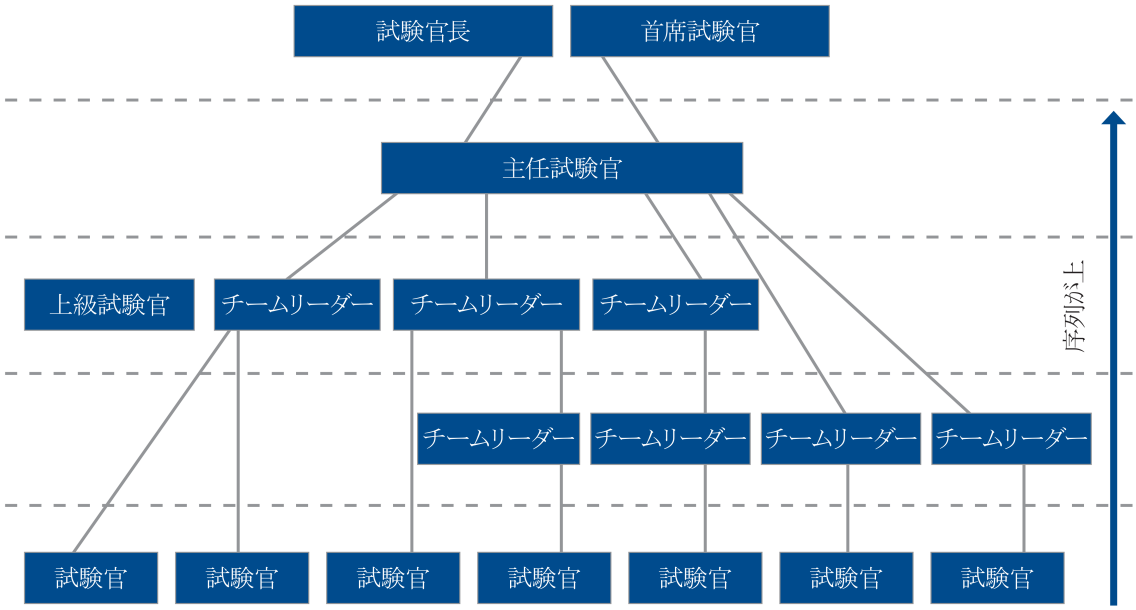
- 主任試験官は、評価要素に付与すべき評点の最終決定者であり、その基準が確定基準です。他のすべての試験官がこの確定基準に従う必要があります。
- IB は、基準の設定に関わった試験官の方が、品質モデルやチームリーダーから基準を学んだ試験官よりも、主任試験官の基準をよりよく理解していると想定しています。
- 「上級」とは、試験官の序列において主任試験官に近いことを意味し、特定の人物が試験官を務めた期間の長さや関連する教育経験の年数を指すものではありません。

IB の採点の原則によれば、生徒の成果物に付与される正しい評点を最終的に決定するのは主任試験官です。IB は、評点の「正しさ」に関して、それぞれに合理的な数多くの見解が存在する可能性があることを認識していますが、成果物を担当した試験官によって生徒の評点が変わってしまう場合、それは公平とはいえません。したがって、主任試験官が基準を設定し、他のすべての試験官がそれに従うことが求められます。

主任試験官は、上級試験官チームのサポートを受けて基準を設定します。IB では、上級試験官は基準について主任試験官と話し合う機会が与えられるため、基準を最もよく理解することになるだろうと考えています。ただし、主任試験官の基準と照らし合わせた採点の信頼性を、すべての試験官についてセッションごとに記録し、各試験官の信頼性の指標を示すようにしています。

試験官には序列があります。履修者数が特に多い科目については、図 23 の例に示すような構造となる可能性があります。

図 23
履修者が非常に多い科目の試験官の序列



序列は柔軟に調整できます。認定用スクリプトまたはシードスクリプトのデータにより、ある試験官が序列上位の試験官よりも主任試験官の基準をよりよく理解していることが示された場合、この点が考慮されます。チームリーダーは基準を適用するだけでなく、説明もできなければなりません。そのため、チームリーダーを指名する際には信頼性以外の要素も検討します。

すべての試験官が同じ基準で採点し、序列に基づいて決定を下す必要がまったくないという状況が理想ですが、実際には、IB がどの評点を付与するかを決定するうえで序列が重要な役割を果たします。特に難しい、または議論の余地のある事例については、答案を主任試験官に確認してもらい、適切な評点を決定します。

評価の完全性

- IB の評価が公平となるには、基準に到達するための平等な機会がすべての生徒に与えられる必要があります。
- 学校による不正または過失や、学問的誠実性に関する方針に違反する生徒の行為は、規則に従った生徒に不利益をもたらします。IB は、このような行為を防止するためにあらゆる措置を講じます。
- 各プログラムの『評価の手順』の「一般規則」のセクションおよび IB 資料『学問的誠実性に関する方針』に記載されている IB の規則は、生徒の学問的不正行為および学校による不正または過失の可能性をできる限り排除するように設計されています。最終的に、学問的不正行為が許容されず、報告されるような学習文化を作り出すことができるのは学校だけです。
- 特定の形式の評価は他の評価よりも学問的不正行為の影響を受けにくいものの、評価を設計する際には、依然として構成の関連性（つまり、実際に評価したい内容をテストすること）を最優先に考える必要があるというのが IB の原則です。
- 生徒の学問的不正行為や学校による不正または過失が疑われる場合は、必ず IB に報告しなければなりません。

学問的誠実性の定義

学問的誠実性とは、教育において指針となる原則であるとともに、「他者からの信頼を勝ちとるために責任ある行動をとる」という、私たちの選択でもあります。また、学習、指導、評価に取り組む際や、正当かつ真正、さらに誠実な学術的成果物を作成する際に、倫理的な意思決定および行動の基礎となります。学問的誠実性は、一連の厳格な規則として押しつけられるべきではなく、むしろ学校内、そして保護者を含むより広いコミュニティーにおいて前向きな文化となるべきです。それが、全員にとってより公平な評価結果に自然につながっていきます。

評価が妥当性をもつためには、1 人の生徒の到達度を、その評価を受けたすべての生徒との比較において、正確に反映しなければなりません。IB はこの理由から、試験の採点、成績付与、バイアスの排除に対して、一貫したアプローチがとれるよう細心の注意を払っています。IB が定める規則や規定は、この平等な機会の創出および公平性の確保のもう 1 つの要素です。

『学問的誠実性に関する方針』では、学問的不正行為を、少なくとも 1 つの評価要素において、その生徒またはその他の者に不当な優位性を与える可能性のある故意または過失行為と定義しています。このような活動は資格の妥当性を低下させるため、その影響は関係する生徒だけでなく、評価を受けた全員に及びます。したがって、IB は学問的不正行為を

重く受け止めています。その予防措置および対応についての詳細は、『学問的誠実性に関する方針』に記載されています。

評価にはその種類によって監視しやすいものとそうでないものがあるため、評価の設計は学問的不正行為を防止するための重要なツールです。例えば、試験では、内部評価のための成果物に比べて他人の力を借りることが難しくなります。IB は評価へのアプローチを設定する際にこのような点を考慮にいれますが、原則として、学問的不正行為を防止するために構成の関連性を犠牲にしてはならないと考えています。

管理のしやすさもまた妥当性の重要な要素であり、学問的不正行為の防止にあたって検討する必要があります。試験会場の設営は、特にデジタル評価においては学校にとって大きな課題となる場合があり、この点も考慮に入れなければなりません。IB は、規則や規定を設定する際に学校の体験を念頭に置くよう努めており、コーディネーターや校長から、どのような実践が役に立ちどのような実践が難しいかについて意見を聞きたいと考えています。

学問的誠実性、および倫理的な行動の意味について詳しくは、「Academic Integrity in the IB: Making the Right Choices（IB における学問的誠実性：正しい選択をする）」（動画）を参照してください。

利益相反

試験内容へのアクセスは IB の組織内で慎重に統制され、IB は利益相反にあたることがないよう、試験を受ける生徒とのつながりを積極的に管理しています。利益相反にあたる可能性が見受けられた場合、スタッフの職務を再編成します。

IB は原則として、試験官が自分の生徒の成果物、ならびに試験官にとって利益相反となる学校（例えば、その試験官が別の科目の個人指導を行っている学校や、その試験官が最近まで勤務していた学校など）の生徒の成果物を採点できないようにしています。非常にまれな状況ではありますが、IB の力が及ばない何らかの理由によってこの原則を守ることができない場合（例えば、求められる基準で採点する資格のある試験官が他に誰もいない場合）には、独立性をもつ 2 人目の試験官が採点の確認を行います。IB は、その試験官の受け持ちの生徒に、他とは違う基準が適用された兆候がないことを確認する必要があります。

学校による不正または過失、および生徒の不正行為を管理する

各プログラムの『評価の手順』の「一般規則」のセクション、および IB 資料『試験実施要項』、または、IB 資料『The conduct of IB Middle Years Programme on-screen examinations』に、不正行為の可能性をできる限り排除するための規則や指示が記載されています。紙ベースの試験については、IB 資料『機密情報である IB 試験資材の安全な保管』に、IB の機密情報である資材の校内保管に関する指示と、IB の方針を実践するために学校が講じるべき手段が記載されています。

試験は、IB 資料『試験実施要項』(DP および CP)、または IB 資料『The conduct of IB Middle Years Programme on-screen examinations』(MYP) に記載された指示に従って、監督されなければなりません。試験監督者は、学問的不正行為が行われないよう、試験時間全体を通して慎重に監督を続ける必要があります。

IB は、評価セッション中に予告なく試験会場を訪問し、必須とされる試験実施要項がすべて守られていることを確認します。これは、学校による IB プロセスの実施状況を確認するための抜き打ち検査です。ただし、不正や過失を防止する責任の大部分は、日々の学校生活の中で高い基準が守られていることを確認できる立場にある校長およびプログラムコーディネーターに委ねられます。さらに、校長やコーディネーターは、教師や生徒たちがベストプラクティスに従い、高いレベルの誠実性を発揮しやすいような学校文化になっていることを確認する必要があります。

試験中の学問的誠実性の基準を守ることに加えて、授業内（および自宅）で取り組む成果物が、本当に生徒本人によるものであることを確認し、剽窃が起こらないようにすることも大切です。

生徒の成果物を「サポート」するウェブサイトや AI ツールが数多く存在するなかで、この種の学問的不正行為に対する最善の防御となるのが教師です。教師は生徒とともに学習に取り組む立場にあり、生徒が提出した成果物がその生徒の普段の水準を反映していない場合、その齟齬に気づくことができます。この理由から、生徒本人が成果物に取り組んだことを、生徒と教師の両方が確認することが求められます。剽窃チェックソフトウェアに頼るだけでは十分ではなく、教師は内部評価の執筆に取り組む生徒にサポートを提供しながら、生徒本人が取り組んだ成果物であることを確認する必要があります。成果物の提出後に剽窃が確認された場合、IB は、代わりの成果物を受領する義務を負いません。

剽窃はその程度にかかわらず一切許容されません。他の著作物からの引用および AI ソフトウェアの使用については、IB 資料『効果的な引用と文献表記』および IB 資料『学問的誠実性に関する方針』の記載に従って、適切に出典を明記する必要があります。IB は、生徒の学問的不正行為の可能性を検知するためにさまざまなソフトウェアを使用します。エビデンスが見つかった場合、正式な調査が開始されます。

学問的誠実性の文化を確立・維持することはすべての IB ワールドスクールに課される要件であり、複数回にわたって違反行為が確認された場合、IB ワールドスクールとしての認定に影響が及びます。

国際的な試験と時間帯がもたらす課題

IB では、大多数の国のシステムとは異なり、すべての生徒が同時に試験を受けることができません。これは、IB ワールドスクールがもつ国際的な性質がもたらす IB に固有の課題です。つまり、一部の生徒は他の生徒よりも先に試験を終えることになり、IB の生徒と教師には、この状況を自分の利益のために利用しない誠実さが求められます。生徒数が一定の水準を超える場合、別々の「試験時間帯」を設けて、1 つの時間帯で生徒が試験を終えるタイミングと、別の時間帯で生徒が試験を始めるタイミングをなるべく近づけるよう

な措置がとられますが、それでも試験の開始時間には大幅なばらつきが生じることがあります。

IB では、このリスクを最小限に抑えるために、試験の内容が共有される可能性のある各種ウェブサイトやソーシャルメディアのプラットフォームを細かくモニタリングするなどの措置を講じています。試験の内容を試験会場から持ち出すことができないという規則は、試験問題がオンラインで共有される可能性を制限するものです。試験時間が 2 時間未満であり、かつ、その日の午前または午後に実施される唯一の試験である場合、午前または午後の試験の開始時間から少なくとも 2 時間は生徒を監督下に置かなければならないという監督規則も、禁止されている情報共有の機会を減らすうえで役立ちます。

また IB は、知識の想起ではなく理解をテストするような試験問題を作成します。つまり、事前に問題を知ることにそれほど大きな価値はありません。最後の措置として、試験後 24 時間は試験の内容について話をしないよう生徒に求めています。これにより、「たわいのない会話」から情報が漏れてしまうことを防ぎます。

IB は、時差による不正行為のリスクを軽減できるよう、引き続き革新を続けていきます。

デジタル評価のメリット

デジタル評価は、学問的不正行為を管理するうえで多くのメリットをもたらします。試験内容にアクセスできる時間は大幅に制限されており、IB はいつ誰が試験内容にアクセスしたかをモニタリングできます。

さらに、学校が試験に加えた変更（時間の追加や一時停止など）も記録され、学校は必要に応じてその根拠を示す必要があります。つまり IB は、公平な実践を確保しながら、妥当と思われる変更を実施する責任を学校に委ねられるようになります。

デジタル評価では、メタデータを使用して試験をめぐる生徒の体験をより詳細に把握することも可能になり、これは学問的不正行為の疑義がある場合に、調査を支えるエビデンスとして使用できます。また、デジタル答案は電子形式のため、疑念が生じるほど類似した解答がないかを一括でチェックすることができます。手書きの解答ではこのようなチェックを行うことはできません。

IB は、デジタル評価が新たな形態の学問的不正行為、特にハッキング行為の対象となる可能性があることを認識しており、この対応策としてさまざまなツールやプロセスを導入しています。

リソース

学問的不正行為の抑制および阻止に関して最も大きな力をもっているのは、教師、学校、生徒自身です。学問的不正行為を容認せず、発生した不正行為に対して速やかに対処するような学校文化を醸成することで、不正行為に対抗することができます。このトピックについては、本資料の「倫理的な考え方の育成」のセクションでさらに詳しく説明しています。『学問的誠実性に関する方針』など、学問的誠実性の文化を築くうえで学校の指針となる資料は、IB ウェブサイトの「学問的誠実性」のセクションにまとめられています。

さらなるサポートが必要な場合、学校で対処されていない学問的不正行為について懸念がある場合、または学問的誠実性に関する新たな問題について意見がある場合は、IB アンサーに問い合わせてください。

全員にとっての公正さを優先する

- 妥当性を維持するためには、形成的評価および総括的評価において学習者の多様性を考慮する必要があり、特定の生徒に影響があってはなりません。IB の評価は、最初からこの学習者の多様性を考慮に入れて設計されています。
- 公平性を確保する最善の方法は、誰もがアクセスできるように評価を設計することです。IB の評価はユニバーサルデザインの原則に従うことで、妥当性に影響を与える可能性のある障壁を排除しています。学校は、筆記試験、口述試験、デジタル試験の実施条件がどの生徒にも有利または不利にならないようにするために、IB が定めたガイドラインに従う必要があります。
- IIB 資料『学習支援と多様な生徒の受け入れに関する方針』は、問題用紙への対応など、追加的な措置や配慮が必要となる状況について説明しています。
- 試験の実施や評価の妥当性に影響を与える可能性のある予期せぬ事態が発生した場合には、IB 資料『特別な事情に関する方針』に定めるとおり、その悪影響を軽減するための一連の措置が講じられます。
- 学習と指導の段階、および形成的評価は、総括的評価や IB の評価に備えて調整する必要があります。受験上の配慮の提供をめぐる重要な原則の 1 つに、その措置が学習、指導、評価において（生徒の通常の作業方法として）使用されているものでなければならないという原則があることを踏まえると、この調整は特に重要な意味をもちます。またこの調整は、デジタル評価を含むすべての評価媒体に適用されるべきです。
- このような配慮の目的は、すべての生徒にとっての公平性を確保することです。判断を下すにあたり、IB は、どうすればすべての生徒にとって公平となるかを検討する必要があります。

すべての生徒ができるかぎり公平な条件の下で、自分の能力を示すことができなければなりません。標準的な評価条件では、一部の生徒が不利な立場に置かれ、到達度を実証できなくなる可能性があります。生徒の力ではどうしようもない特別な事情が、そのパフォーマンスに影響を与える場合もあります。公平性を確保するための措置を講じる際は、この 2 つの要因を考慮に入れる必要があります。

評価における公平性を確保する最善の方法は、すべての生徒が同じ条件下で同じ評価を受けるようにすることです。これを実現するため、評価は生徒がもつ多様な要件を慎重に検討しながら開発されます。学習者の多様性に対処し生徒のニーズを満たす最も効果的なアプローチは、学びのユニバーサルデザインと評価のユニバーサルデザインの原則に基づいて評価を設計することです。これらの原則に従って評価を設計し、次のステップとし

て、『学習支援と多様な生徒の受け入れに関する方針』で説明されている受験上の配慮を通してアクセシビリティーを確保します。

受験上の配慮は、障壁を取り除き、すべての生徒が試験を受けられるようにするものです。これらの障壁は、ニューロダイバーシティー（神経多様性）、障害、困難、社会情緒をめぐるさまざまな状況が評価において適切に考慮、または配慮されていない場合に発生する可能性があります。

特定の生徒の独自のニーズに関して、妥当と思われるあらゆる調整が検討対象となります。詳細は、『学習支援と多様な生徒の受け入れに関する方針』を参照してください。

特別な事情とは、生徒がコントロールできない理由により、生徒のパフォーマンスに悪影響が及ぼされるかもしれない状況のことです。『特別な事情に関する方針』に、この影響への対応策および緩和策が説明されています。受験上の配慮は、障壁を取り除き、すべての生徒が評価を受けられるようにするために実施されます。このような配慮は、障壁に直面する生徒を有利な立場に立たせるものではなく、むしろ公平なアクセスを可能にし、すべての学習者が他の生徒と同等に自分のスキル、知識、能力を発揮できるようにします。

受験上の配慮に関する原則

受験上の配慮に関する原則は、『学習支援と多様な生徒の受け入れに関する方針』に記載されています。評価では、各プログラムのねらいに反映される、青少年の発達段階を考慮する必要があります。受験上の配慮は、形成的評価と総括的評価の両方において、また学習と指導全体を通じて、すべてのプログラムで検討されなければなりません。

IB は、すべての科目で生徒に付与される成績が、生徒の到達度を正確に反映したものであることを確認しなければなりません。これを実現するため、受験上の配慮の有無を問わず、すべての生徒に対して同一の評価基準が適用されます。

妥当な調整を含む受験上の配慮の申請は、生徒が評価に関する支援を受けるために、事前に行われなければなりません。口述試験、筆記試験、デジタル試験のいずれの場合も、遡及的に申請することはできません。

生徒のために申請する受験上の配慮は、いかなる評価要素においても、生徒を有利な立場に立たせるものであってはなりません。

受験上の配慮は、すべての評価要件を満たす素質のある生徒に提供することを想定しています。

受験上の配慮が必要とされる場合、学校は、『学習支援と多様な生徒の受け入れに関する方針』に明記されている IB の規準に基づいて配慮を提供することができます。申請する受験上の配慮は、その生徒の普段の学習方法を反映したものでなければなりません。学習プログラム全体を通して、生徒のニーズを最も効果的に満たすような配慮を選択し、総括的評価への円滑な移行を図ることは学校の責任です。生徒の通常の学習方法とは異なる受験上の配慮が申請された場合、非常に特別なケースに限り、例外的に承認されます。

IB は、国際的な視野に基づく教育理念を掲げています。受験上の配慮に関する要件をもつ生徒も含めた公平性を達成するため、関連する方針が、さまざまな国や地域（統治領）で受け入れられている実践方法を考慮した結果として策定されています。

受験上の配慮の申請は、それぞれの理由に基づいて個別に判断されます。受験上の配慮が IB や他の機関によって過去に承認されたものであっても、そのことが申請中の配慮の承認をめぐる決定に影響することはありません。

IB は、児童生徒に関するすべての情報を機密情報として取り扱います。必要な場合のみ、適切な IB の担当者と IB の資格授与委員会（Final Award Committee）のメンバーに情報が共有されますが、すべての関係者がこの情報を機密情報として取り扱うよう指示されます。

受験上の配慮を提供するにあたって、IB が定める条件を学校が満たしていない場合、または、学校が IB の許可なく受験上の配慮を提供した場合、当該科目およびレベルにおいて成績が付与されない可能性があります。

生徒が試験での使用言語に堪能でない場合、それがあらかじめ特定された学習のサポートの必要性から生じている性質のものであることが証明されれば、受験上の配慮が許可されることがあります。

試験官が、生徒の状況や受験上の配慮を知ることは禁止されています。

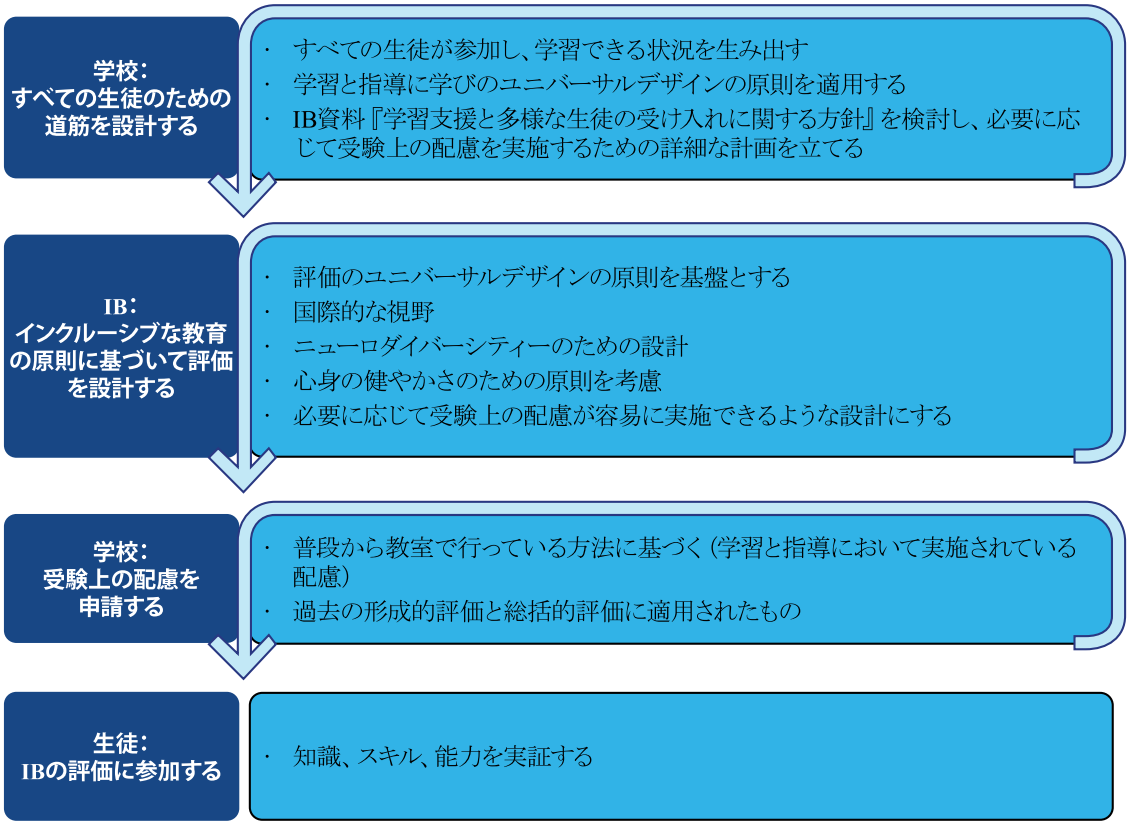
同様に、教師が評価課題の内部評価を行う際に、採点に斟酌を加えてはなりません。

実施可能な受験上の配慮の一覧は定期的に見直されます。学校が提案する代替案については、同様の事情のある他の生徒にも共通して適用できる可能性がある場合に限り、検討されます。

生徒自身、または承認された受験上の配慮が試験中の他の生徒の妨げとなる場合、その生徒は別室で受験しなければなりません。またその場合、その生徒は各プログラムの試験の実施要綱に従って監督されなければなりません。

試験中、受験上の配慮に関連して生じた問題、または生徒が直面した予期せぬ困難な事態に関しては、すべて「IB アンサー」に速やかに報告するようにしてください。

図 24
受験上の配慮を実施する



評価の免除

いかなる評価要素においても、通常、評価が免除されることはありません。ただし、評価要素またはその一部が、生理学上、生徒によって遂行できない場合は、評価を免除することが許可されます。評価要素について評価の免除を申請する前に、他のあらゆる合理的な調整手段を慎重に検討するようにしてください。評価の免除は、妥当な根拠がある場合にのみ許可されます。評価要素を生徒が遂行できない理由は、書面による明確かつ十分な裏づけがなければなりません。

評価の免除に関する原則やプロセスについての詳細は、『学習支援と多様な生徒の受け入れに関する方針』に記載されているインクルーシブな配慮のリストを参照してください。

デジタル評価がもたらす受験上の配慮の機会

デジタル評価では、評価が表示される方法を生徒が細かくコントロールできます。コンピューターデバイスでは、個人の好みやニーズに合わせてさまざまなフォント、文字サイズ、色を提供でき、これにより、すべての生徒が調整を加えることが可能になります。ユニバーサルデザインの原則を実施することでアクセシビリティーは向上するものの、依然として受験上の配慮を提供しなければならないケースもあります。これを実現するうえで、デジタル機能は、このような受験上の配慮に代わる手段を提供できると考えられてい

ます。利用可能な受験上の配慮は、『学習支援と多様な生徒の受け入れに関する方針』に記載されています。

デジタル試験のための受験上の配慮の目的は、紙ベースの試験と同様、障壁を取り除くことです。公平性を確保するために、承認済みの受験上の配慮は、デジタル試験と紙ベースの試験で一貫していなければなりません。

特別な事情

特別な事情または不測の事態とは、生徒のせいではない事情または事態で、かつ、そのパフォーマンスに悪影響を及ぼす可能性があるものとして定義されます。これには、一時的な疾病や怪我、重度のストレス、特段に困難な家庭の事情、忌引き、生徒の健康または安全を脅かし得る事象が含まれます。また、暴動や自然災害など学校コミュニティー全体に影響する事象も含まれます。

学校側の過失は、特別な事情に含まれません。教職員の問題を含め、すべての生徒が確実にプログラムおよび評価の要件に従うようにするのは、学校の責任です。

特別な事情に含まれるものと含まれないもの、およびその緩和措置に関する詳細は、『特別な事情に関する方針』で確認することができます。

評価のユニバーサルデザイン

これまでのセクションで、IB 評価に対する公平なアクセスを確保するために、受験上の配慮が必要となるケースがどのように管理されるかを説明しました。ただし、理想的な解決策は、最初の段階から障壁のない評価を設計することです。評価のユニバーサルデザインでは、アクセシビリティー、インクルージョン、平等、文化的多様性、感受性への配慮に重点が置かれます。このアプローチに従うことで、IB の評価は、設計、開発、実施の段階で多くの障壁が評価モデル全体から取り除かれるように設定されています。

よりインクルーシブで構成の関連性が高い評価を最初から作成することは、受験上の配慮を導入・実施する必要性を最小限に留めることにつながり、これは多くの生徒にメリットをもたらします。

評価のユニバーサルデザインは、すべての学習者を受け入れる学習環境の創出を目指す、学びのユニバーサルデザインという幅広い枠組みの 1 つの側面です。このアプローチは、ニューロダイバーシティー、好み、能力、課題、関心、文化、言語、背景などを含むさまざまな多様性を考慮に入れます。

学びのユニバーサルデザインの原則は、カリキュラムと評価の設計、カリキュラムと評価の開発、学校の管理など、教育のさまざまな側面に適用されます。IB の学びのユニバーサルデザインについての詳細は、 ラオら（Rao et al., 2016）を参照してください。

評価のライフサイクル

- 評価を作成するプロセスは、各段階がその前の段階に支えられ、次の段階へとつながっていく循環型サイクルとして考えるべきです。
- 試験（および対応するマークスキーム）の作成には平均で 18 か月かかるため、IB は複数のセッションの試験を同時並行的に作成します。

評価プロセス全体を、作成、生徒が受験するセッション、試験官による採点、そして結果発表へとつながるサイクルとして考えることができます。プロセスの重要な側面は、IB が各セッションから学び、次のセッションのために評価の質を改善するという点にあります。

図 25 は、プロセスを説明する 1 つの方法にすぎません。各段階を細分化したり、他の方法で組み合わせたりすることができますが、評価のライフサイクルを説明する効果的な方法として示されています。表 6 は、各段階の要約と、関連するセクションとのつながりを示すものです。

図 25
評価のサイクル

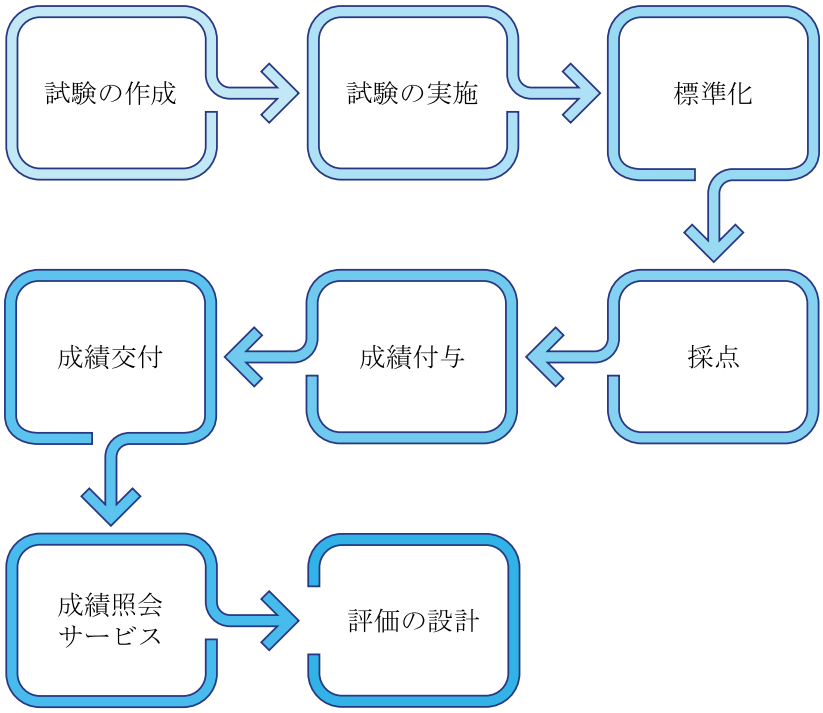


表 6
評価のライフサイクルの各段階

評価サイクルの段階	説明
試験の作成	個別の試験を作成するプロセス。これには、出題するトピックの決定から、具体的な問題の作成と編集、他の言語への翻訳、適切な形式での構成、そして最後に実施される必須の品質チェックまでのすべてのステップが含まれます。
試験の実施	生徒が各学校で評価を受ける段階。
標準化	上級試験官が試験官チームに生徒の成果物の採点方法を説明し、採点の品質保証プロセスのために、評点確定済みの答案を特定するプロセスです。
採点	試験官が個々の生徒の成果物を確認し、付与する評点を決定します。主任試験官が定めた指示に従う必要があり、採点が正しく行われているかどうかが定期的に確認されます。
成績付与	上級試験官が、評点（試験ごとに異なる）をどのように成績（常に同じ意味をもつ）に換算するかを決定します。成績付与後の活動には、試験官の成績分布をモニタリングすること、および特定の問題のエビデンスに基づき、誤った成績を付与される恐れのある生徒を把握することが含まれます。
成績交付	成績が学校および生徒に対して開示されます。
成績照会サービス	学校が試験プロセスで誤りがあったと思われる箇所を報告し、生徒の成果物を再検討するよう IB に依頼します。
評価の設計	生徒のパフォーマンスおよび試験問題の解釈は、評価の設計に有益な情報をもたらします。IB はこの情報を使って、評価すべき対象と、IB がとるアプローチを改善します。これには、課題の数や種類が含まれます。

評価のライフサイクルについての詳細は、「The life cycle of an exam: The steps that we go through from an exam being created to grades being awarded（試験のライフサイクル：試験の作成から成績付与までのステップ）」（動画）を参照してください。

評価サイクルに対するデジタル評価の影響

デジタル評価の導入によって評価サイクルの原則が変わることはありません。デジタル評価は、例えば、紙ベースの試験や答案をスキャンセンターに送付する必要がなくなるなど、一部の段階の時間短縮につながりますが、評価ライフサイクルの各段階を完了しなければならないということは変わりません。

評価の作成

- 評価の作成においては、生徒が評価を受けられるようになる前に、いくつかの最終プロセスを完了する必要があります。これには、以下が含まれます。
  - 評価の作成と内容のチェック
  - 組体裁と校正
  - 使いやすさ
  - 翻訳
  - 学校への評価の送付
- 生徒によって申請された調整（受験上の配慮）も、この最終プロセスで考慮されます。
- 紙ベースの試験として実施される場合も、デジタル試験として実施される場合も、評価作成サイクルの原則は変わりません。

優れた評価要素

IB はカリキュラム開発者と作成チームが、「セクション A：評価の原則」で説明されている優れた評価の条件に加えて、以下の重要な要素を念頭に置くことを期待しています。

- 評価とは、生徒ができないことを特定するのではなく、できることを実証する機会を生徒に与えるべきものです。
- できる限り幅広い生徒が評価を受けられるようにする必要がある一方、妥当な個別最適化を可能にする必要があります。
- 最も優れた評価とは、すべての生徒が利用できる評価です。IB は、評価の作成プロセス全体を通してユニバーサルデザインを採用することで、受験上の配慮の必要性をできる限り減らしています。
- 評価は、科目の『指導の手引き』に記載されたカリキュラムのみをテストするものでなければなりません。
- どのように採点されるかを念頭に置きながら評価を作成することにより、評価で確認しようとしていることに対して評点が付与されるようにすることができます。つまり、マークスキームが試験問題と同時に作成されます。また、試験官に対する予想許容差についてのガイダンスもここに含まれます。

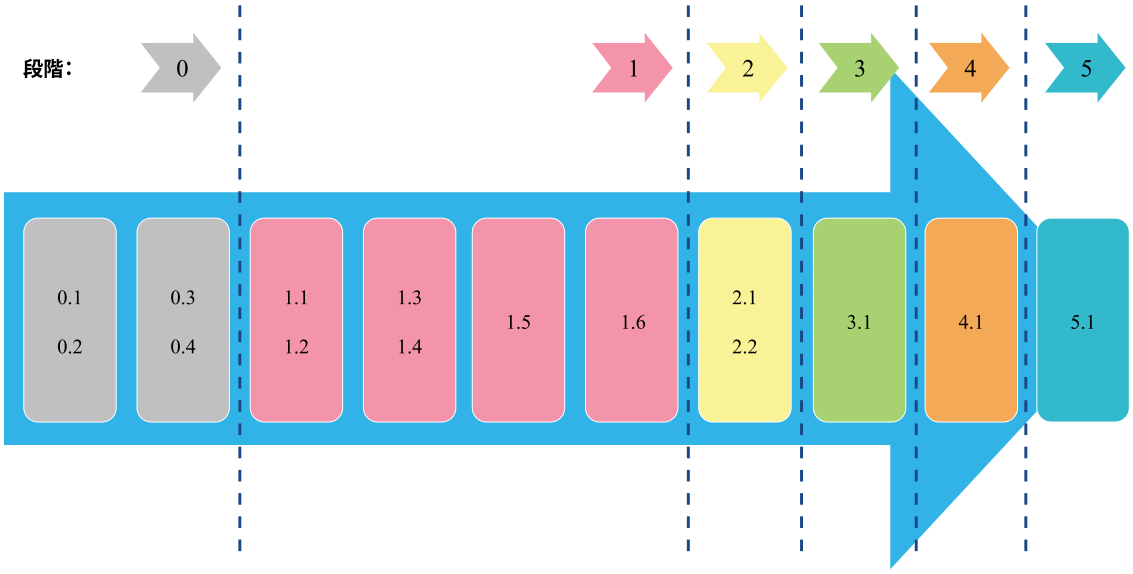
指示用語

指示用語とは、シラバスや試験問題で用いられる主要な用語や語句で、特定の指示に対して、生徒が何をしなければならないかを示すものです。また、生徒に期待される解答の種類や深度も示しています。

指示用語に固有の課題として、他の言語に翻訳される過程でそのわずかなニュアンスが失われる可能性があるということが挙げられます。これを防ぐため、IB では各使用言語の『指導の手引き』に指示用語を掲載し、その意味を説明しています。翻訳された試験問題の確認作業において、正しい指示用語が使用されていることを確かめることも重要な責任です。

試験制作の概要

図 26
試験作成プロセス



段階0:
リソース承認
0.1 リソースの計画
0.2 リソースの評価と選定
0.3 立ち上げ会議
0.4 日程と役割の決定

段階1:
内容の承認
1.1 評価の草稿作成および修正
1.2 相互評価
1.3 評価編集会議
1.4 外部アドバイザー向けの草稿作成
1.5 外部アドバイザーの報告および修正
1.6 内容の承認

段階2:
体裁の承認
2.1 評価の構成
2.2 校正

段階3:
使いやすさの承認
3.1 チェッカーの報告

段階4:
翻訳の承認
4.1 翻訳と修正

段階5:
5.1 受験者への提供

図 26 は、試験作成プロセスに含まれる各段階を示すものです。評価の作成は、約 18 か月もの長期にわたるプロセスであり、作成者、サブジェクトマネージャー、科目の専門家、編集者、デザイナー、外部のチェッカーや翻訳者といった、さまざまな外部の専門家が関与します。

マークスキームを試験問題と一緒に作成することは非常に重要です。これにより、評価で確認しようとしていることは何かという観点から、問題と採点を完全に整合させることができます。

内容の承認までのプロセス

プロセスの最初の段階は、試験問題の作成者を選ぶことです。通常は主任試験官に依頼しますが、IB は、学校とのつながりや教職員研修のワークショップの運営などといった、利益相反を生み出す可能性のあるあらゆる要因を検討します。

試験の作成に関わる役割として、以下が挙げられます。

- 作成者（外部）
- 外部アドバイザー、標準化担当者（外部）
- チェッカー、該当言語に精通した話者（外部）
- サブジェクトマネージャー（IB スタッフ）
- 制作編集者（IB スタッフ）
- コピーエディター（IB スタッフ）
- デザイナー（IB スタッフ）
- 翻訳者（外部）
- 言語校閲者（外部）

複数の時間帯で使われる試験問題を作成する場合には、完全に独立した試験作成プロセスが実施されます。

作成と内容の承認

作成者の仕事は、試験問題とマークスキームの初稿を作成することです。この初稿を、他の作成者およびサブジェクトマネージャーが複数回にわたってレビューします。一部の科目では、その後、正式な評価編集会議が開かれ、そこで各試験について以下の点から確認を行います。

- 問題の認知的要求度、および試験の全体的な難易度の範囲
- バイアスの有無
- アクセシビリティー
- カリキュラムとの適合性
- 試験問題が公開された評価モデルと一致していることの確認
- 過去の試験問題およびサンプル資料との関係
- 作成された問題が、公開済み資料のどの試験問題見本とも一致していないことの確認

この会議によって完成した新たな草稿を、外部アドバイザーまたは標準化担当者がレビューします。

試験問題とマークスキームの草案作成に関与していない科目の専門家が外部アドバイザーまたは標準化担当者を務め、新たな視点を提供するとともに、受験対象のすべての試験の全体的な難易度、過去の試験との比較、カリキュラムとの適合性などについてコメントします。外部アドバイザーが提案した変更点を作成者とサブジェクトマネージャーが確認し、それを基に最終稿を作成します。

最終稿は次に内部のコンテンツレビューにかけられ、コピーエディターとデザイナー、および必要に応じてサブジェクトマネージャーと作成者が、専門的なチェックを行います。

最終的な試験問題の内容の承認は作成者が行い、IB サブジェクトマネージャーが、IB 独自の品質基準に従って試験問題が適切に精査されたことを確認します。

マークスキームはこの時点では最終承認されず、生徒の答案に照らした標準化プロセスにおいて再度確認されます。

組体裁と校正

内容の最終承認が完了すると、試験問題の文言が固定されます。デザイナーは、最終版の設問を受験用の形式（紙ベースまたはデジタル）に合わせて配置し、体裁を整えます。

従来の紙ベースの試験では、指示（またはルーブリック）を記載した表紙を追加し、適切なスタイルを適用し、バーコードや空白ページを追加して、問題冊子を作成します。このプロセスには、求められる質に合わせて図を再描画したり、適切な画像をオンラインで入手したりといった作業も含まれます。

デジタル評価では、デザイナーが適切な開発環境で試験問題を作成する必要があります。これには、指示やルーブリック、適切なスタイルなどが含まれます。

どちらの場合でも、制作編集者が、完成した評価を最終承認された内容と照らし合わせて校正作業を行います。

使いやすさの承認

チェッカーは、この時点までで評価の作成プロセスに関与していない科目の専門家が務めます。解答にかかる時間を考慮に入れながら、生徒になったつもりで試験を受けます。このチェックの目的は、間違いを見つけること、および指示や問題に曖昧な点がないことを確認することです。

チェッカーは、試験を受け終わった後でマークスキームも確認します。

受験者が少ない言語の科目では、文学の専門家ではなく、その言語をある程度流暢に操る話者がこのチェック作業を行います。

チェッカーによる正式なフィードバックを検討した後、評価の最終版が完成し、安全な方法で学校に送付されます。

品質管理

試験問題の間違いや曖昧さは、試験を受ける生徒に大きな影響を与える可能性があるため、そのような問題がないように試験を作成することが非常に重要です。IB ではこの点を非常に真剣に捉えています。

正式な最終承認プロセスに加えて、評価作成サイクルの各段階で品質チェックのプロセスが設けられています。

翻訳

IB は、生徒がさまざまな使用言語で試験を受けられるようにしています（言語の試験はのぞく）。現在、大部分の試験が英語、フランス語、スペイン語で提供されています。地域の合意に基づき、一部の科目は、ドイツ語、日本語、韓国語でも提供されています。

試験を翻訳しなければならないということは、初稿の作成段階から考慮されており、IB の作成者およびコピーエディターは、翻訳をめぐって発生し得る問題を十分に認識しています。正式な翻訳プロセスは、英語の試験問題の完成後に開始されます。

特定の言語で受験することが生徒にとって有利にも不利にもならないためには、翻訳のプロセスによって問題の意味が変わらないことがきわめて重要です。IB の外部翻訳者は、

評価における専門用語の正確性を確保するために雇用された科目の専門家です。翻訳完了後、バイリンガルの外部校閲者が翻訳文を元の試験問題（問題文だけでなく試験全体）と比較して、最終的な品質チェックを行います。

調整済み試験問題

理想的には、すべての生徒が問題なく受験できるように評価を設計すべきです。ただし、ユニバーサルデザインを採用しただけでは対応できず、試験問題への調整が必要となるような申請や要件も存在します。このような調整は、『学習支援と多様な生徒の受け入れに関する方針』に従って管理され、その作成と学校への送付には、調整なしの試験問題と同じ方法がとられます。

試験

- 試験の目的は、すべての生徒に同一の体験を提供し、学問的不正行為の可能性を最小限に抑えるよう管理された環境で、知識、理解、スキルの応用を実証する機会を生徒に与えることです。
- 試験セッションを設定する際、IB はすべての生徒のニーズと学校の管理のしやすさをバランスよく両立し、生徒の期待に応えるためにできるだけ早く採点を完了する必要があります。
- IB は、学問的不正行為の可能性を最小限に抑えるため、試験会場における行動について明確な規則を定めています。この規則は、最新のテクノロジーおよび環境の変化に合わせて更新されます。
- 試験日程は、多くの相反する優先事項を考慮した妥協案であり、世界的に見て最も悪影響が少ない選択肢を表しています。
- 予期せぬ出来事、受験上の配慮、特別な事情、日程変更への対応に関する詳細は、該当するプログラムの IB 資料『評価の手順』に記載されています。

試験期間中の生徒は、大きなプレッシャーとストレスにさらされます。IB は、試験期間に対する包括的な原則として、以下のような措置によってストレスの軽減を図っています。

- 試験セッションの長さを制限する。
- 可能な限り、他の試験と日程が重複しないようにする（例えば、全国試験など）。

これらの措置と、生徒のパフォーマンスについて妥当な結論を導き出すために十分な評価を実施する必要性、および学問的不正行為を防止する必要性とを、適切なバランスで調整する必要があります。

試験会場の準備と管理

学校は、該当するプログラムの IB 資料『試験実施要項』に定められた厳格な規則に従って試験を実施しなければなりません。

セキュリティーを確保しながら各試験を適切に実施することは、きわめて重要です。学問的不正行為があったという認識を学校や関係者がもった場合、生徒の成績の価値が大きく低下することになるためです。本資料の「評価の完全性」で説明しているように、IB は、国際的な信頼性を維持するために数多くの方針や手順を策定しています。

学校や生徒がコントロールできない何らかの理由により、IB が生徒の成果物を採点できなくなるという状況は、必然的に起こり得るものです。評点を推定することにより生徒への影響緩和を試みることはできますが、この措置は、裏づけとなる他のエビデンスが存在して初めて可能となります。紙ベースの試験でこの措置を実行するために、IB では、それぞれの解答用紙（答案）を別の日にスキャンセンターに送付することを義務づけています。これにより、すべての答案が輸送中に紛失する可能性を減らしています。

デジタル評価については、評価のために提出された成果物のコピーがアップロードプロセスの各段階で入手できるため、DP と CP においてこのリスクが軽減されます。

試験日程の作成

IB は、セッションの長さを最小限に抑えるという目的のため、試験日程において異なる科目を同じ時間枠に割り当てています。生徒が履修できる科目の組み合わせを最大限考慮していますが、試験時間が重複するケースがわずかながらに発生することを認識しています。この場合の対応については、該当するプログラムの『評価の手順』に記載されています。

試験日程は、1 日の試験時間ができる限り 6 時間半を超えないように設定されています。1 つのコースに関連する複数の評価要素の試験日程を組む際は、通常、午後から次の日の午前中にかけて 2 つまたは 3 つの試験を連続で実施する、というパターンを採用します。常に実現できるとは限らないものの、この方が、1 つのコースのすべての試験を同日に実施するよりも好ましいとされています。

ほとんどの生徒にとって、これにより試験が 2 週間（MYP）および 3 週間（DP と CP）の期間にわたって均等に分散されることになり、特定の試験で十分な成果を出せなかったと感じても、夜の間に気持ちを切り替えることが可能になります。

試験日程の設定の原則

以下の箇条書きは、試験日程の作成の基礎となる原則をおおまかにまとめたものです。

常にすべての原則を満たせるとは限らず、そのような場合には妥協案が必要になります。IB は、試験実施の少なくとも 1 年前に試験日程を公開します。

- IB のプログラムが提供されている国と地域（統治領）の数からみて、公休日、祝日または学校の休暇（5 月 1 日は除く）を考慮することは不可能です。
- 中東の学校は木曜日と金曜日が週末にあたるため、この 2 日間は試験を実施しないことが望ましいものの、現時点ではそのような対応をとることはできません。
- 特定の地域または文化とつながりをもつ科目については、その点を考慮に入れるように努めます。例えば、アラビア語の文学や言語の試験は金曜日には実施されません。

- IB は、科目の組み合わせに関する登録データを利用して、科目日程の重複により影響を受ける生徒の数が世界的に最小限になるようにしています。
- また、生徒が 1 日に 2 つの異なる外国語（使用言語ではない）の試験を受けることがないように配慮されています。
- 複数の試験があるコースでは、予期しない出来事がすべての評価要素に影響し、生徒にとって不利に働くというリスクをできる限り抑えるため、少なくとも 2 日間にわたって試験日程が組まれます。
- 生徒が 1 つの科目の復習に短いスパンで集中して取り組めるよう、試験が 2 日以上にわたって実施される場合には、連続する 2 日間でその科目の試験が受けられるような日程が組まれます。
- 短い期間に一気に試験を受けさせるのではなく、試験期間全体にわたって試験を分散させる必要があります。つまり、可能な場合には、最も受験者数の多い科目（「英語」「歴史」「数学」など）の試験が何日も連続しないように日程が組まれます。
- これと同じ理由で、「言語」と「理科」の試験が 2 週間（MYP）および 3 週間（DP と CP）の試験期間の各日に実施されるよう、試験日程の調整を試みます。
- IB 内部の試験処理要件により、通常、特定の科目（履修者が多い科目）の試験が日程の早い段階で実施されることになります。

予期しない出来事への対応

各プログラムの『評価の手順』に、さまざまな予期しない出来事への対応方法が記載されています。判断がつかない場合は、IB アンサーに問い合わせてください。適切なアドバイスを提供します。

受験上の配慮と特別な事情

このトピックについては、本資料の「全員にとっての公正さを優先する」のセクションで詳しく説明しています。

IB は、すべての生徒ができるかぎり公平な条件の下で、自分の能力を示すことができなければならないと考えます。ただし、標準的な評価条件では、学習のサポートを必要とする生徒が不利な立場に置かれ、到達度を実証できなくなる可能性があります。受験上の配慮はこのような状況において承認されるもので、受験の際の特別な措置、または試験問題への調整という形で実施されることがあります。

試験セッション中に発生した出来事のうち、生徒のせいではなく、かつ、生徒のパフォーマンスに悪影響を及ぼす可能性がある出来事は、特別な事情と呼ばれます。これには、一時的な疾病や怪我、重度のストレス、特段に困難な家庭の事情、忌引き、生徒の健康または安全を脅かし得る事象が含まれます。また、暴動や自然災害など学校コミュニティー全体に影響する事象も含まれます。学校側の過失は、特別な事情に含まれません。

これらの措置についての詳細は、各プログラムの『評価の手順』を参照してください。

試験の日時変更

試験の日時変更は、試験プロセスの完全性にとって大きなリスクとなります。生徒が他の生徒よりも前に、または後に試験を受けることになり、学問的不正行為が起こりやすくなります。試験日の前日に試験の日時を変更することは、いかなる場合においても認められません。これは、その試験を受ける大多数の生徒に影響を与える違反行為のリスクが過度に高まるためです。

IB が定めた試験日程とは異なる時間、または異なる日に 1 つ以上の試験を受験することが認められる状況、および試験の日時変更を申請するためのプロセスは、プログラム・リソース・センターに収載されている各プログラムの『評価の手順』および IB 資料『試験の準備に関する方針』に詳しく定義されています。

採点

- 効果的な採点とは、一貫性があり、正確であり、かつ成果物の質を反映した点数を生徒に付与することを意味します。
- 主任試験官によって検討された見解を正とし、すべての試験官が採点においてこの見解を再現する必要があります。
- 評点と成績は同じではありません。難易度の低い試験では評点が上がる可能性がありますが、付与される成績は同じになるはずです。

評点と成績

生徒の総括的評価を実施・活用するうえでの重要な要素として、成果物を採点することと、成果物に成績を付与することの違いを理解する必要があります。

- 採点において、生徒は、マークスキームなどの枠組みに照らしてその成果物に適した点数が与えられます。これは、評価課題において生徒が正解した度合いを表すものです。評点そのものにそれ以外の意味はありません。
- 成績を決める際は、試験官は生徒の成果物の質を確定済みの基準に照らして判断します。この基準では、課題の難易度および課題を完了した割合が考慮に入れられます。したがって、成績にはある程度の意味または関連性があり、通常は、他の評価のパフォーマンスとの同等性を確保することをねらいとします。

非常に難しい問題のほんの一部に正解することによって高い成績を示すことができる場合があり、それと同様に、難易度の低い問題に数多く正解しても、同じ成績を示すことができない場合があります。

後のセクションで説明するように、成績によって示される基準は、必ずしも生徒が達成した内容を参照して説明されなければならないわけではありません。IB ではこのアプローチを採用していますが、他にも高い一貫性をもつ認知度の高いシステムとして、他の生徒との相対的なパフォーマンスに基づく基準を使ったものも存在します。

IB の評価では通常、評点は全体のパフォーマンスの指標として使われます。そして、同じ評点を付与された生徒のパフォーマンスを検討し、その評点よりも高い評点を獲得した生徒に特定の成績を付与する境界線として、区分点（成績換算）を決定します。このプロセスは、「成績の付与と集約」のセクションでさらに詳しく説明されています。

採点に対するアプローチ

IB ではさまざまな評価の手段が使用されるため、各評価に対して適切な採点のアプローチを選ぶことが非常に重要です。採点のアプローチは、カリキュラム開発プロセスにおいて決定されます。

分析的マークスキーム

分析的マークスキームが使用される評価では、生徒の解答の幅が狭くなることが予想されます。

このマークスキームは、設問に対する総合点を解答の異なる部分に対しどのように配分するかについて、試験官に詳細な指示を与えることを目的としています。非常に具体的な解答が予想される構造化された設問であっても、マークスキームは、生徒がとるであろうさまざまなアプローチとよくある間違いを考慮に入れて、試験官が一貫した採点を行うための十分な情報を提供する必要があります。想定外の解答や解答例にはない正答の配点について、試験官が専門的な判断を下さなければならない場面が必ず生じますが、マークスキームは、その判断を下すために必要なガイダンスをできる限り提供する必要があります。

設問において生徒が正答または誤答する部分は、必ずしも予想されたパターンにあてはまるとは限りません。これは特に、解答の最初の部分の間違いが後の部分に影響を与える可能性がある、構造化された長文問題に見られる問題です。IB は、このような問題を設計する際に、設問の最初の部分を間違えてしまうと、それによって残りの部分を解くことができなくなるという状況が起こらないよう留意しています。たとえそうであっても、分析的マークスキームでは、特定の種類の誤答をどのように採点し、設問の一部で間違えてしまった生徒の解答を最終的にどのように扱うかについて明示的なガイダンスを提供する必要があります。

評価規準

詳細なマークスキームが適切ではない評価では、代わりに評価規準が適用されます。

評価規準は、解答の細かい内容よりも、生徒が実証すべきパフォーマンスの種類に焦点をあてたものです。実証される専門知識のさまざまな段階が、レベルの説明に反映されています。

規準のレベルの説明を適用するにあたっては、「ベストフィット」モデルが使用されます。評価規準を適用する試験官は、採点対象の成果物を全体的に見て、それに最もあてはまる到達度を選択する必要があります。ある到達度を選ぶにあたって、そのレベルのすべての要素が満たされている必要はありません。また、各規準の最高レベルが、完璧な成果物を表しているわけでもありません。

評価規準を使用して採点される評価要素の多くで、1 つの規準ではなく複数の規準が使用されます。これらの規準が互いに独立していることが重要です。解答における 1 つの要素について複数の箇所で評点を与えられるという状況は公平ではないためです。

通常、生徒に出題される設問は毎年変わりますが、使用される評価規準は変わりません。これは、評価される対象は変わらないという基本的な性質によるものです。問題が変わる部分は、通常、規準に明示的に示されていない具体的な詳細に関連しています。多くの場合、主任試験官が提供するマーキングノートに、各設問に評価規準を適用する際のガイダンスと、関連性の高い詳細情報の例が示されます。

包括的な規準：マークバンド

1つの成果物を採点する際に、それぞれの評価規準を分けることが適切でない場合もあります。これは通常、それぞれ独立した個別の規準を設けることが不可能な場合に起こります。このような場合には、個別の規準ではなくマークバンドが使われます。マークバンドは、基本的に包括的な規準を示したものです。合理的な評価領域に沿って生徒のパフォーマンスの差異を明確にすることが要件となっていることから、マークバンドのレベルの説明は評点範囲ごとに分けられます。

表7

IBのマークバンドの例

評点	レベルの説明
0	成果物が、以下のレベルの説明に記されたいずれの基準にも達していない。
1〜3	・ 問題についての理解がほとんど示されていない。科目固有の用語が使われていない、またはその使い方が一貫して不適切である。 ・ 問題について最低限の説明しかなされていない。論点が表面的で、しばしば不明確である。 ・ 答案が説明的である。分析が表面的、またはまとまりがない。さまざまな視点についてまったく触れていない、またはわずかしか触れていない。結論が示されている場合、それが非常に表面的である、または答案の他の部分と整合していない。
4〜6	・ 問題についての基本的な理解が示されている。科目固有の用語は使われているが、不適切な使用が多い。 ・ 問題の説明が基本的で、十分に発展されていない。論点が不正確または不明瞭なことが多く、答案が伝えようとしていることがしばしば不明確である。 ・ 分析が限定的であり、答案全体が分析的ではなく説明的である。さまざまな視点について限定的にしか議論されていない。短絡的な結論が含まれている。
7〜9	・ 問題についてある程度の理解が示されている。科目固有の用語が、時おり適切に使用されている。 ・ 問題について十分な説明が成されているが、その説明は一部で明瞭性と発展性に欠けている。関連性の高い論点が述べられているが、正確さや詳細さに欠ける。 ・ 答案に分析が含まれるが、その分析は十分に発展されていない。さまざまな視点についてある程度議論されている。結論が含まれている。
10〜12	・ 問題についての深い理解が示されている。科目固有の用語が、おおむね適切に使用されている。

評点	レベルの説明
	・ 問題について説明は明瞭だが、さらなる発展が必要とされる。述べられた論点は関連性が高く正確だが、詳細さに欠ける。 ・ 答案に批判的な分析が含まれるが、その分析は十分に発展されていない。さまざまな視点について議論されている。答案が 1 つの結論に帰結し、その結論は提示された議論と一貫性がある。
13〜15	・ 問題についての非常に深い理解が示されている。科目固有の用語を適切かつ正確に使用している。 ・ 問題の説明は明瞭で、効果的に発展されている。論点は関連性が高く、正確で詳細である。 ・ 答案に、効果的に発展させた批判的分析が含まれている。さまざまな視点に関して、批判的な議論がなされている。答案が理路整然とした明確な結論に帰結し、その結論は提示された議論と一貫性がある。

レベルの説明は、生徒の成果物に見られるであろうさまざまな特徴を網羅するため、かなり長くなる傾向があります。またこれらの説明は、コースの目標に直接立ち返るものです。評価規準とあわせて、ベストフィットのアプローチが適用されます。試験官は、生徒の成果物がレベルの説明にどの程度適合しているかに基づき、そのレベルの説明に対応する範囲の中から、どの評点を付与するかを判断する必要があります。例えば、あるマークバンドの範囲が 7 点〜9 点だったとします。試験官は、生徒の成果物がそのマークバンドのレベルの説明にどの程度適合しているかに基づいて、特定の評点を付与します。

自動採点

数は多くないものの、多肢選択問題で構成される評価要素は自動採点されます。自動採点は、テクノロジーを使用して、事前に定義されたマークスキームに照らして生徒の課題を評価するプロセスです。これらの多肢選択問題において、マークスキームは、解答が客観的に正しいか間違っているかを規定しています。このような評価要素のマークスキームはきわめて明確であり、試験官が自ら判断を下す必要はありません。

試験官の判断を必要とする採点

IB の評価は、一部自動採点が可能なものもありますが、そのほとんどが分析的マークスキームまたは採点規準を必要とするものであるため、試験官による採点が求められます。

試験官は、「採点ツール」と呼ばれるソフトウェアを使って、生徒の成果物を電子的に採点します。これにより、視聴覚素材を含む生徒の成果物の電子コピーに対して、チェックマークを入れる、点数をつける、メモを残すなどの作業が可能になります。さらに、パフォーマンスの録画や口述試験など、エビデンスとして提出できる成果物の幅が大きく広がります。

IB の採点について詳しくは、「The two methods of marking: How marking is carried out by examiners and teachers（2 つの採点方法：試験官と教師は採点をどのように実施するのか）」の動画を参照してください。

標準化

標準化のプロセスには、評価手順のいくつかの要素が含まれます。標準化の目的は、信頼性の高い採点に向けて試験官の準備を整えることです。

信頼性の高い結果を得るためには、各答案をどの試験官が採点した場合でも、内容の質に基づいて同じ評点が付与されるようにすることが重要です。この点から、試験官が IB の試験セッションでライブマーキングを行うためには、主任試験官がその評価要素について定めた採点基準への理解を示すとともに、その基準に沿った採点ができなければなりません。

標準化の最初の手順は、採点のための適切なスキルと経験をもつ試験官を採用することです。採用された試験官は、採点の原則と実践を理解するための研修を受けます。

試験が実施されたら、主任試験官の主導のもと、上級試験官チームが会議を開いてマークスキームについて話し合い、答案サンプルを確認します。この標準化会議のねらいは、試験官が目にすると予測されるさまざまな解答を、マークスキームが十分に網羅していることを確認することです。マークスキームによって、生徒が導き出す解答のすべてを網羅することはほぼ不可能ですが、主任試験官と上級試験官チームが会議の中で答案サンプルを確認することで、各評点で求められる解答の種類に関する試験官向けの明確な指示とガイダンスを、それまでに確認された解答に基づいて作成できるようになります。

標準化会議の結果として、主任試験官によって確定採点された一連のスクリプト（答案）が得られます。このスクリプトを使って他の試験官の研修を行うことで、全員が主任試験官と同じ基準で採点できるようにします。主任試験官の採点基準に沿って採点できることを実証した試験官のみが、実際の試験セッションで答案を採点することが認められます。またその採点基準は、厳格な品質保証手順によって採点期間全体を通して定期的に確認されます。

品質モデル

どの試験官が生徒の成果物を採点するかにかかわらず、主任試験官が採点した場合と同じ結果を得られるようにすることが重要です。

評価が実施され、生徒の解答がいくつか提出された後、主任試験官は他の上級試験官と標準化会議を開催します。上級試験官たちによって、複数の解答に対する評点がそれぞれ定められ、それが品質モデルの設定に使用されます。確定採点済みのスクリプトには、練習用スクリプト、認定用スクリプト、シードスクリプトの 3 つのカテゴリーがあります。品質モデルはこの 3 つのカテゴリーから構成されます。

練習

練習用スクリプトの目的は、主任試験官が設定した採点基準について試験官たちが学ぶための資料を提供することです。典型的な答案の採点方法を示すとともに、マークスキームに関して固有の理解が必要になる一般的な状況を特定します。

低い評点から高い評点まで、さまざまな生徒のパフォーマンスを幅広く含むものが優れた練習用スクリプトといえます。これにより試験官は、どこに評点を付与すべきか、また、どのような解答が良い解答（または悪い解答）といえるのかを見極められるようになります。

マークスキームや採点規準、および標準化会議で主任試験官が作成したマーキングノートを確認した試験官は、練習段階に入ることができます。練習用スクリプトは、主任試験官の採点基準を示すものです。試験官は練習用スクリプトを採点しながら、必要に応じてチームリーダーに指導を求めることができます。

練習用スクリプトをすべて完了した試験官は、主任試験官の採点基準を理解し、確信をもって適用できるようになっているはずです。その後、試験官は資格認定プロセスに入ります。

認定

認定用スクリプトの目的は、正しい基準に沿って採点できることを示す機会を試験官に提供することです。無事に認定を終えた試験官は、実際の試験セッションでの採点に取り組めるようになります。

通常、5 つの認定用スクリプトが 1 セットとして提供されます。試験官が認定プロセスにおいて主任試験官と同じ評点を付与しなかった場合、与えるべき評点とその理由についてのフィードバックが提供されます。その後、試験官は、採点基準を理解したことを改めて示すため、別の認定用スクリプトを使って再度認定プロセスに取り組みます。

優れた認定用スクリプトとは、低い評点から高い評点まで、さまざまな生徒のパフォーマンスを幅広く含むものです。また、マークスキームに追加された具体的な指示を理解していなければ正しく採点できない答案が含まれている必要があります。このような具体的な指示は、主任試験官や上級試験官チームの予想に基づき、採点に関するさまざまな課題に対応するヒントとなるものです。これは、標準化会議での答案確認作業の結果に基づいて提供されます。認定用スクリプトは、試験官を脱落させる意図をもって選定されるのではなく、試験官が採点中によく見かけることになる解答の種類を示すものです。

この段階では、主任試験官の採点基準を理解し、それを評価用に提出された成果物に適用できることを示すチャンスが、各試験官に 2 回与えられます。試験官は、必ず認定用セットに合格しなければなりません。

各認定用セットは、複数の確定採点済みのスクリプトから構成されます。認定セットに取り組んでいる間は、スクリプトについてチームリーダーと話をすることはできません。第 1 回目の認定セットの結果が許容範囲内に入っていた試験官は、ライブマーキングに進みます。

試験官の採点が許容範囲から外れていた場合は、主任試験官の基準に沿った採点をする方法について、チームリーダーからフィードバックを受け取ります。フィードバックについて振り返った後、採点基準への理解を示す二度目のチャンスを与えられます。

第 2 回目の認定セットの採点が許容範囲内に入っていた試験官は、ライブマーキングに進むことができます。

ライブマーキングとシード

シードスクリプトは、品質保証プロセスの一部です。シードスクリプトは、試験官のパフォーマンスをモニタリングし、正しい基準に従って採点していることを確認するために使用されます。

時間が経つにつれて、試験官の採点が主任試験官が設定した基準から離れ始めることがあります。この理由から、IB は、主任試験官によって確定採点されたシードスクリプトを使って試験官の採点を定期的に確認します。シードスクリプトの見た目は他のスクリプトと同じであり、品質保証目的のスクリプトを採点していることが試験官にわからないようになっています。通常、試験官は、10 件に 1 件の割合でランダムにシードスクリプトを採点することになります。

試験官がシードスクリプトに対して主任試験官と同じ評点を付与した場合、採点をそのまま続行することができます。一方、試験官が不正確な評点を付与し、求められる基準で採点していないことが示された場合、IB が対応措置を講じます。

始めに、試験官が再び基準に合わせて採点できるようにするため、上級試験官がフィードバックとガイダンスを提供します。そのようなサポートがあってもなお基準に沿って採点できない場合、その試験官が採点作業を続けることは認められません。

各評価要素には採点許容差が設けられ、確定済みの基準からのある程度の乖離は認められています。詳しくは、「許容差」のセクションを参照してください。

効果的な標準化の指標

効果的な標準化とは以下をもたらすものです。

- 採点に関する試験官の質問に対応し、採点の一貫性を確保できる意欲の高い上級試験官チーム
- 付与された評点の根拠をはっきりと説明する、明確で曖昧さのない指示（マークスキームやマーキングノートとして提供）
- 研修や品質保証の目的で使用される、確定採点された練習用スクリプト、認定用スクリプト、シードスクリプト

許容差

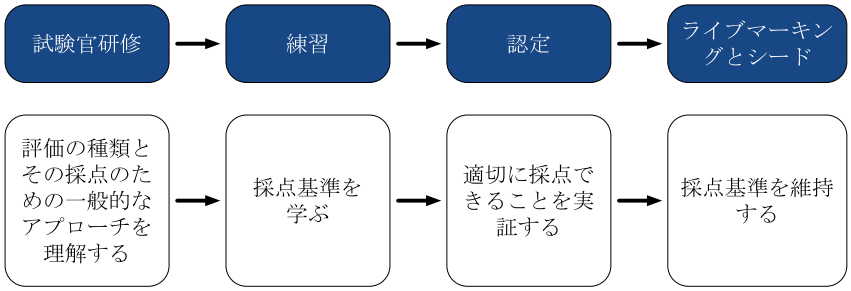
IB は、採点が簡単な問題ではなく、重要なことを確認するための問題を出題することを信念としています。答えが正しいか間違っているかの 2 通りしかない種類の問題については、試験官が主任試験官とまったく同じ評点をつけることを期待することは道理にかなっています。その他の問題、特に理解力や分析力を確認するような問題の場合、ある程度の

判断が必要となります。実際の採点においてこれは、問題に対する適切な解答について同じ専門知識と理解を有している 2 人の試験官が、同じ成果物にまったく同じ評点を付与しない場合があるということを意味します。2 人の評点には 1〜2 点の誤差が生じることがあります。

IB が試験官のパフォーマンスを確認する際には、この正当な意見の相違を反映するために許容差という考え方が使用されます。例えば、小論文に対する主任試験官の評点が 46 点で、その問題の許容差が 2 点の場合、44 点から 48 点を付与した試験官はすべて適切な基準に沿って採点していると考えます。

IB は、試験官が答案を採点する際に、その評点が確定評点（主任試験官の評点）にどれほど近くなければならないかを定義しています。この許容される誤差が許容差です。複数のパートや規準から成る問題を確認する場合、IB は全体的な評点の差と、各パートや規準の差の両方をモニタリングします。試験官が適切な基準に沿って採点していることを示すためには、この両方が許容差内に収まっていなければなりません。

図 27
採点の整合



試験官研修

試験官は、評価セッションを開始する前に研修を受けることができます。研修の目的は、評価の種類とその採点のための一般的なアプローチを試験官が理解できるようサポートすることです。将来のセッションの試験に関する詳細は、この研修には含まれません。

判断に迷う答案および通常とは異なる答案

どのように採点すべきか判断できない答案があった場合、試験官はチームリーダーに助言を求めます。助言を得てもなお公正に採点できるという自信がもてない答案は、その試験官よりも上級の試験官に送られ、採点されます。特に問題だとされる答案は、主任試験官に報告され、主任試験官がその答案の採点方法について最終的な判断を下します。

同様に、通常とは異なる答案や、問題用紙に調整が加えられた生徒の答案は、主任試験官によって採点されます。そのような答案にも他の生徒と同じ基準が適用されるように、採点のバランスを慎重に見極めるのは主任試験官の仕事です。

学校のつながり

IB は、試験官が自分の学校の生徒を採点すべきではないという原則を定めています。

これを管理するために、IB は試験官全員に対して生徒や学校とのつながりをすべて開示するよう求め、利益相反となるような学校に在籍する生徒の成果物を受け取らないようにしています。このプロセスについて詳しくは、本資料の「利益相反」のセクションを参照してください。

試験官による注記と注釈

IB の総括的評価の目的は、生徒のパフォーマンスを測定することです。試験官に期待されるのは、求められる基準に照らして生徒の成果物を採点することのみです。IB は、採点の助けになる場合のみ、コメントを残すように試験官に指示しています。

試験官が生徒や教師に向けた形成的なフィードバックを書くには、成果物の適切な評点を決定し、そのうえでどうすればその成果物を改善できるかを説明することが必要です。この作業は優れた指導の中核を成すものであり、決して簡単な作業ではないと IB は認識しています。時間をかけて慎重に取り組まねばならず、一貫した基準に照らして採点を行うという本来の業務から試験官を逸脱させるものです。簡単に言えば、IB が試験官に求めているのは、採点とフィードバックという 2 つの作業を低い水準で行うのではなく、採点という 1 つの作業を高い水準で遂行することです。

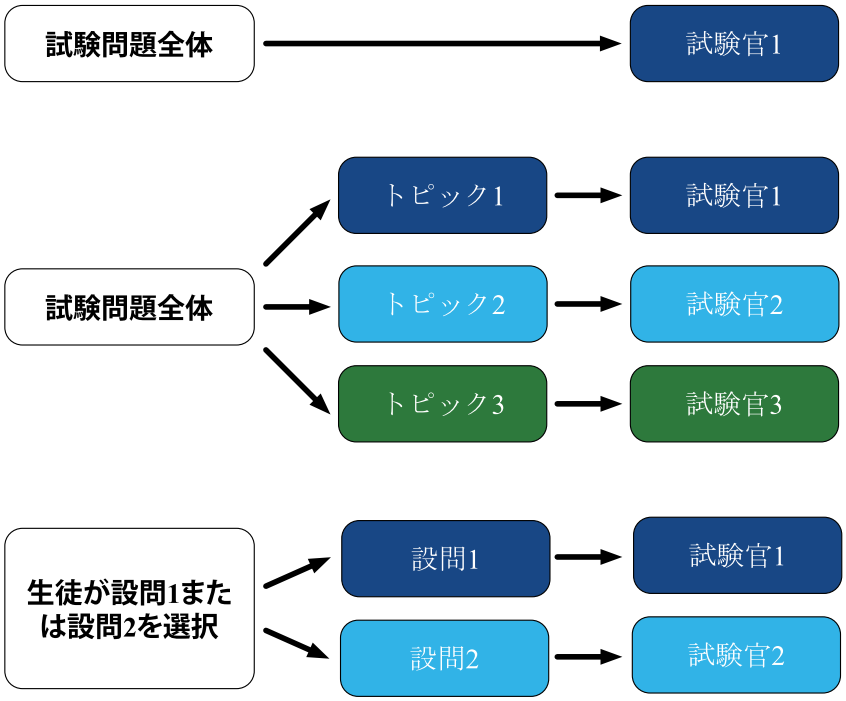
また、試験官が判断を下せるのは、目の前にある 1 つの成果物についてのみである一方、経験豊かな教師は、生徒にフィードバックを提供する際にさまざまな情報を参考にすることができます。したがって、教師ほどの情報をもたないということは、試験官のフィードバックの質にも影響します。

このような理由から、IB は、試験官は適切な基準に従って生徒の成果物を採点することのみに集中し、生徒や教師にフィードバックを提供するためにコメントを残すべきではないというスタンスを明確に示しています。

採点プロセスを円滑に進めるため、試験官は評点が付与された部分を明確に示す必要があり、曖昧さが残る場合にはその説明のために適切なコメントを追加することが求められます。これは、基準が守られていることを確認し、評点が付与された箇所について学校に透明性を提供するという点で、IB をサポートするものです。

設問項目グループ

図 28
試験問題を設問項目グループに分ける



設問項目グループは、試験官が 1 人の生徒の答案の設問すべてを採点し、次の生徒の答案に移ってまた一から採点を始めるよりも、同じ設問を繰り返し採点する方が負担が少ないという考えに基づいています。

設問項目グループごとに練習用スクリプト、認定用スクリプト、シードスクリプトが作られ、試験官は各質問項目グループについて、求められる基準に合わせて採点できることを示す必要があります。難しそうに思えるかもしれませんが、これは、試験官が 1 つの設問の基準を理解できない場合でも、その他の設問は採点できることを意味します。つまり、試験のすべての設問を必要な基準で採点できないために、採点業務に参加できないということがなくなります。

集約

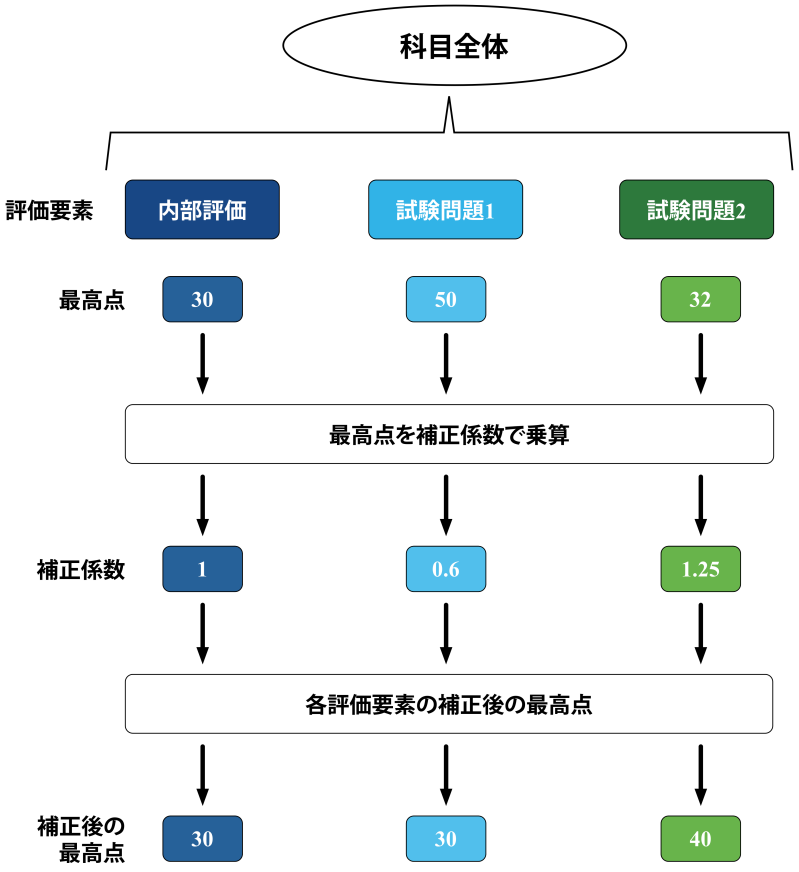
- 集約とは、評価要素を合わせて全体的な結果を生成するプロセスのことです。
- 各評価要素が適切な比率で最終的な評点に寄与するには（配点比率）、要素の評点の調整が必要となる場合があります。
- IB は、評点が低い設問や要素を評点が高い設問や要素で補うことができる「補完モデル」を採用しています。
- 生徒の最終的な科目の成績は、各評価要素の成績ではなく、各評価要素の評点を集約した結果で決まります。

集約は、さまざまな評価要素の評点（および区分）を合計して、最終的な評点または全体的な成績区分を決定するプロセスです。このためには、全体的な評価要素の評点（または区分）の補正が必要となる場合があります。

補正は、コース全体の評価に対する各評価要素の貢献度という観点から、各評価要素の望ましい配点比率を維持するために実施されます。これは、各評価要素の評点が、その科目の全体的な合計点に対して適正な比率となるよう、評価要素の評点を乗算または除算することを意味します。その評価要素に対して設定された成績区分に対しても、同じことがあてはまります。成績区分は当初、その評価要素の最高評点を基に決定されます。

各評価要素の補正済みの最高点は、その科目の『指導の手引き』で定められた評価要素の配点比率と同じになります。

図 29
3 つの評価要素をもつ科目における補正の例



配点比率の概念は、最終結果への貢献という点において、IB が評価対象の要素に与える相対的な重要度を反映したものです。例えば、ある評価要素が主にデータまたは情報源の解釈を確認するものであり、その配点比率が 30%の場合、これは、コースの他の目標に対して、解釈の重要性が約 30%を占めることを意味します。多くの場合、複数の評価要素が同様の目標をもつため、この計算にはあまり意味がありません。

配点比率にはもう 1 つ、重要な側面があります。それは、それぞれの評点を合わせた点数を杓子定規に全体合計点にするのではなく、各課題と採点規準に適した評価の合計点数を IB が設定できることです。そしてこの合計点数を集約して、最終的な成績を導き出すことができます。

評価における集約は、さまざまな評価要素の評点を合計し、生徒の最終成績を決定するプロセスです。このアプローチにより、評価要素の配点比率から導き出される所定の合計点に縛られることなく、課題の難易度と規準、および評価の設計意図に基づいて評点を設定できる柔軟性がもたらされます。つまり、評価要素の配点比率も考慮事項の 1 つではあるものの、例えば評価や項目の相対的な難易度が思わぬ傾向を見せることにより、実際の結果が意図された結果と異なる可能性があるということです。この IB のアプローチには、個々の評点を所定の配点比率に細かくそろえるという作業は含まれません。最も重要視されるのはあくまでも全体的な結果です。「補完モデル」として知られるこの方法では、評点が低かった部分を評点が高かった他の部分で補うことが可能になり、個々の要素の評点ではなく合計点に重点が置かれます。したがって最終的な科目の成績は、各評価要素の成績の平均としてではなく、すべての評価要素の評点の合計から導き出されます。そのため、評価要素単位で同じ成績をとった生徒が、最終的には異なる成績を付与されるということもあり得ます。

モデレーション

- モデレーションは、生徒の成果物を再採点するプロセスではなく、教師の採点基準を確認するためのプロセスです。
- 望ましいモデレーションとは、ある生徒の内部評価と、地球の反対側にある学校に通う生徒の内部評価に、同じ評点が付与されることを意味します。IB ではこれを「グローバル基準」と呼んでいます。
- モデレーションでは、生徒の成果物の質だけではなく、その評点を与えた理由についての教師の説明もエビデンスとして使用されます。

モデレーションにはダイナミックサンプリングが使用され、DP および CP の内部評価、そして MYP の e ポートフォリオを対象として実施されます。

モデレーションを定義する

多くのケースにおいて、IB が評価したいと考える生徒の資質は、制限時間のある正式な試験では確認することができません。このような場合、最も妥当なアプローチは、内部評価としてこのような資質をテストするよう教師に依頼することです。このアプローチが適切な理由については、「授業内評価と内部評価の役割」のセクションで説明されています。

この方法は有意義な結果につながりますが、その一方で、それぞれの教師が採点基準について異なる解釈をするというリスクも生み出します。2 つの学校の 2 人の教師が、同じ成果物に対して別の評点を付与する可能性があります。IB では、試験官全員が求められる基準について共通の理解を得られるよう研修とテストを行っていますが、これを教師全員に実施することは現実的ではありません。

IB では、IB の教師がすべての生徒の成果物を同じ基準で採点すると確信しており、IB がすべきことは、グローバル基準に合わせて教師の採点に適宜調整を加えることのみだと考えています。IB ではこのために、教師の採点サンプルを提出してもらい、主任試験官の基準と比較しています。この比較によって提供されるデータから、必要に応じて各教師の評点を調整する数式を導き出します。2 つの評点群を統計的に比較し、必要だと判断された場合には、その教師が（その評価要素に対して）学校のすべての生徒に付与した評点に調整が加えられます。教師の採点が一貫して基準よりも低い、または高い場合は、その教師の評点に一律で同じ調整が加えられますが、評点範囲の上部または下部で採点基準が低くなっている、または高くなっている場合は、教師の評点範囲全体で複数の異なる調整が加えられることがあります。

重要な注意点

- モデレーションの目的は、生徒の成果物を採点する際に、教師が評価規準をどの程度正確かつ一貫して適用しているかを確認することです。
- モデレーションの結果として、学校の採点結果が上下に変動することもあれば、そのまま変更されないこともあります。
- モデレーション係数が意味するのは、教師の採点の質が低いということではなく、あくまでもグローバル基準に沿っていないということです。

実践的な理由から、IB は、個々の教師ではなく学校に対してモデレーションを行います。そのため、学校のすべての教師が同じ基準で採点をすることが非常に重要です。モデレーション係数はすべての教師の採点結果に適用されるため、適切な調整が加えられるようにするためには、学校の採点結果を正しく表すサンプルを選ぶ必要があります。

コーディネーターがモデレーションの対象となる内部評価の採点結果を確認する際は、以下を検討するとよいでしょう。

- 学校内のすべての科目担当教師（内部評価の採点を担当する教師）が同じ基準で採点しているか。
- すべての生徒を同じ基準に沿って一貫性をもって採点しているか。
- 特定の評点を付与した理由について教師が明確な説明を提供しているか。

評価規準を正確に提供するための教師向けのガイダンスは、本資料の「学校へのフィードバック」のセクション、および MYP、DP、CP の現行の『評価の手順』の資料を参照してください。

モデレーション係数の計算に関する技術的な詳細は、本資料の付録「内部評価のモデレーション：詳細」を参照してください。

モデレーションと教師に対する期待事項

評価のために提出された成果物は最初に教師が内部で採点し、IB が外部モデレーションを行います。

教師は、関連する科目の『指導の手引き』を読み、内部評価の規準の詳細を確認する必要があります。各評価規準には、学習成果物が特定のレベルに到達している場合にその成果物に見られる特徴を記述した「レベルの説明」と、それに対応する評点が示されています。「レベルの説明」には主に、「何を達成できたか」という肯定的な側面が述べられています。ただし、到達度の低いレベルの説明では、達成できなかった点に言及されている場合もあります。

教師が内部評価課題を採点する際は、評価規準の「レベルの説明」を使い、不明点を明らかにしながら判断しなければなりません。規準は、ベストフィットのアプローチを用いて適用する必要があります。ある評価規準で複数のレベルの説明に合致する特徴が成果物に見られる場合は、その規準の到達度を最もバランスよく言い表しているレベルを選択すべきです。つまり、「レベルの説明」に書かれた側面を必ずしもすべて達成していなくても、その評点を与えることが可能だということです。最上位のレベルの説明は、まったく非の打ちどころのないパフォーマンスを意味するわけではありません。

1 つのレベル内に複数の評点の可能性が与えられている場合、成果物が当該レベルの説明に含まれている特徴をより広く満たしている、すなわちもう 1 つ上のレベルに近い到達度であれば、高い方の評点を付与します。説明内容を達成している度合いが小さければ（その下のレベルに近い場合）、低い方の点数をつけます。

評点は整数のみを使用しなければなりません。部分点（分数や小数を用いた点数）は認められません。

規準は個別に検討されるべきです。ある規準で高いレベルに到達した生徒が、他の規準でも高いレベルに到達するとはかぎりません。同様に、ある規準で到達度の低かった生徒が、他の規準でも到達度が低いとはかぎりません。教師はまた、生徒全員の評価結果が特定の分布を示すと考えるべきではありません。

一部の科目においては、レベルの説明で具体的な指示用語が使用されていることもあります。その場合、指示用語は『指導の手引き』の関連セクションに示されているとおりに解釈されなければなりません。

内部評価課題にモデレーション係数が適用されたからといって、教師の採点作業が間違っていたと考えるべきではありません。この調整はあくまでも、合意されたグローバル基準に採点結果を合わせるためのものです。各試験セッション後に発行される各科目の科目レポートに、内部評価のパフォーマンスに関する教師向けの有益な情報が提供されています。IB 資料『教師用参考資料』（各科目）もプログラム・リソース・センターで入手できます。

各学校が評価のために提出した成果物のすべてに、同じモデレーション係数が適用されます。したがって、1 つの学校の中で、同じ科目の内部評価を複数の教師が担当する場合は、内部標準化を行って、学校内の評価規準をできる限りそろえることが重要です。これには、採点に入る前に、評価規準について話し合ったり、評価のために提出された成果物を確認したりといった作業が含まれます。

教師と生徒は、学問的誠実性に関連した概念、特に生徒本人が取り組むことや知的財産などの概念を理解する必要があります。生徒の成果物は、必ず生徒自身によって作成されたものでなければならず、また、各科目の『指導の手引き』に記載された要件に従って準備されていなければなりません。協働は認められていますが、協働と共謀の違いをすべての生徒に明確に説明する必要があります。より詳細な情報については、本資料の「倫理的な考え方の育成」のセクションを参照してください。

この情報の要約版が必要な場合は、本資料の付録「印刷可能な資料」に含まれている「内部評価の採点：モデレーションプロセスにおける教師への期待事項」のポスターを印刷することができます。

生徒の成果物の選定

- すべての生徒に公平な評点を付与する必要があります。したがって、どの生徒の成果物も、モデレーションにおいて教師の基準を示す例として使うことができます。
- 何をもって優れた成果物といえるか、また質の低い成果物といえるかは、教師によって異なる可能性があります。そのため、学校が提出するサンプルに、さまざまな評点の成果物が幅広く含まれるようにすることが重要です。
- 学校間の透明性と公平性を確保するため、IB はすべての成果物について、生徒の最終的な評点に寄与するエビデンスを確認する必要があります。

透明性を確保し学問的不正行為の疑義を排除するため、学校ではなく IB が、モデレーションのサンプルとなる成果物を特定することが大切です。IB は広範なガイドラインを使って、そのガイドラインの制約の中でモデレーション係数の信頼性をできる限り高めるとともに、実際の生徒の選定が無作為に行われるようにしています。

最初の原則は、いかなる学校に対しても、できる限り少ないサンプル数で信頼性の高いモデレーション係数を導き出すことです。これにより、学校の負担が軽減されるとともに IB にかかるコストも抑えられます。IB のコストが上がると、それが試験料金という形で学校に転嫁されることになります。ただし、信頼性の高いモデレーション係数を決定するために教師の採点結果のサンプルを増やす必要があると IB が判断した場合は、学校によって生徒の成果物のサンプル数が変わることもあります。

2 つ目の原則は、モデレーション係数が評点の範囲全体にわたって、すべての生徒に対して公平であると IB が確信できなければならないということです。IB はこれまでの経験から、評点の範囲内の各段階において、教師の期待の高さが異なる可能性があるということを認識しています。そして、質の低い成果物に対してグローバル基準よりも甘い採点をする教師が、質の高い成果物に対してはグローバル基準よりも厳しめに採点することがある

ということも理解しています。この理由から、学校の評点範囲全体を適切に表せるよう、内部評価のサンプルを慎重に選定します。

IB の傾向として、満点を獲得した生徒の成果物をモデレーション用サンプルに選定することはあまりありません。これにより、評点範囲の上の方にいる生徒の成果物が教師によって厳しめに採点されていた場合、モデレーションによって上方修正することができます。また通常は、0 点を付与された生徒が選ばれることもありません。

例外となるケース

モデレーションは、たとえ非常に採点が難しい成果物であったとしても、教師がグローバル基準に合わせて採点をしているかどうかを確認するものです。採点が難しい成果物の場合、教師は評点を付与した理由について、詳しく説明する必要があるかもしれません。例えば、生徒が成果物を完成させるために教師から付加的なサポートを得た場合、教師はそれを採点コメントの中で説明し、試験官がその生徒の成果物に対する教師の採点を確認する際に、この内容を考慮に入れます。

モデレーション係数を導き出せない場合

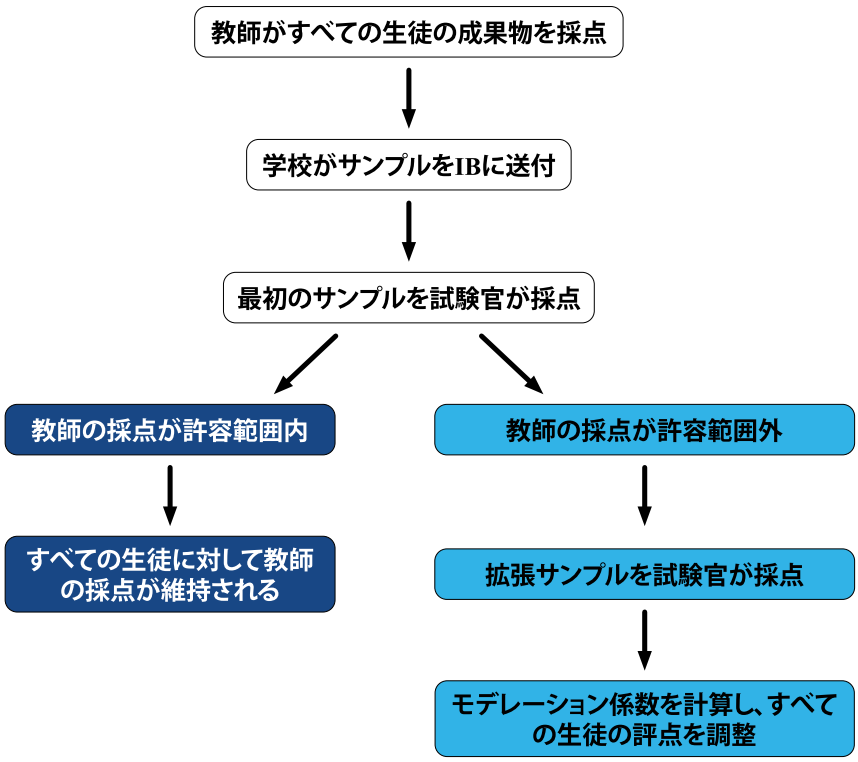
場合によっては、提出された成果物のサンプルからモデレーション係数を導き出せないこともあります。これは、試験官の評点と教師の評点の差にばらつきがある場合や、試験官の評点によって、教師の採点がグローバル基準に比べて甘すぎるまたは厳しすぎることが示された場合に起こります。このような場合 IB は、学校にさらなるサンプルの提出を求め、公平なモデレーション係数が確実に適用されるようにします。すべての生徒の成果物を、試験セッションが完全に終了するまで保管しておかなければならないのはこのためです。非常にまれに、公平なモデレーション係数が適用できない場合は、試験官の評点が付与されます。

ダイナミックサンプリング

- ダイナミックサンプリングを使用したモデレーションとは、教師の採点が許容差内に収まっている場合、IB はその評点を受け入れるということを意味します。
- 教師の評点のいずれかが許容差を超える場合、IB はモデレーション係数を適用します。
- 教師の内部評価の採点を確認するすべての試験官が、品質保証のためにモニタリングされます。

ダイナミックサンプリングを使用したモデレーションの原則として、教師がサンプルを許容差の範囲で採点しており、グローバル基準を理解していることが示された場合、その教師のすべての評点が求められる基準を満たしていると判断されます。許容差を超えていた場合は、モデレーション係数が計算され、適用されます。これが図 30 で説明されています。

図 30
ダイナミックサンプリングを使用したモデレーション



モデレーションが公平であるためには、試験官が主任試験官によって設定されたグローバル基準を理解し、それに沿ってモデレーションを行う必要があります。モデレーションを行う試験官はグローバル基準について研修を受け、その後、この基準に合わせてモデレーションしていることを確認するためのモニタリングを受けます。「標準化」のセクションで説明されているように、主任試験官は確定採点された内部評価の成果物として、以下の 3 種類を用意します。

- 練習用スクリプト（答案）：グローバル基準について説明する
- 認定用スクリプト（答案）：試験官がグローバル基準を理解していることを確認する
- シードスクリプト（答案）：試験官がグローバル基準を維持していることを確認する

1 つの学校のすべての内部評価課題を同じ試験官が確認します。ただし試験官には、さまざまな学校の成果物が順不同で提示されます。これにより、最初に確認した生徒の成果物に基づいて、教師が基準よりも甘く、または厳しく採点しているという意見を固め、そのパターンを同じ学校の残りのサンプルに見出そうとする、という試験官の傾向を防ぐことができます。またこれにより、シードだと気づかれないようにシードスクリプトを入れ込めるようになります。モデレーション係数が必要かどうかが決まった後、学校の採点結果に対して要約コメントを提供するよう試験官に依頼することがあります。

ダイナミックサンプリングの品質モデルは、他の採点方法よりも複雑です。IB はまた、教師による採点が厳しすぎる場合、甘すぎる場合、あるいは適正な場合など、さまざまなケースについて、試験官が自信をもって見直せているかを確認する必要があります。図 31

は、すべてのシナリオを網羅するために品質モデルで必要とされる、内部評価スクリプト（答案）の範囲を示しています。

図 31
品質モデルで必要とされる内部評価スクリプト（答案）の範囲

	状況	
試験官が評点範囲の上端付近で採点	1	教師の採点が厳しすぎる
	2	教師の採点がグローバル基準に沿っている
	3	教師の採点が甘すぎる
試験官が評点範囲の中間付近で採点	4	教師の採点が厳しすぎる
	5	教師の採点がグローバル基準に沿っている
	6	教師の採点が甘すぎる
試験官が評点範囲の下端付近で採点	7	教師の採点が厳しすぎる
	8	教師の採点がグローバル基準に沿っている
	9	教師の採点が甘すぎる

IB によるモデレーションの実施方法について詳しくは、「The two methods of marking: How marking is carried out by examiners and teachers」（動画）をご覧ください。

成績の付与と集約

- どのセッションで試験を受けるかにかかわらず、成績の意味は常に同じである必要があります。
- この点を確保するために、成績付与のプロセスにおいて評点から成績への換算が行われます。
- 成績区分は、受験者群全体の結果や、評価のために提出された成果物に対する専門的な判断など、さまざまなエビデンスを用いて決定されます。
- 最も重要なのは、個々の評価要素ではなく、生徒が受け取る全体的な成績です。

評点と成績は同じものではないということに留意してください。この理由について詳しくは、「評点と成績の違い」のセクションを参照してください。成績付与のプロセスは、特定の事例において、評点と成績の対応関係を決めるための手段です。

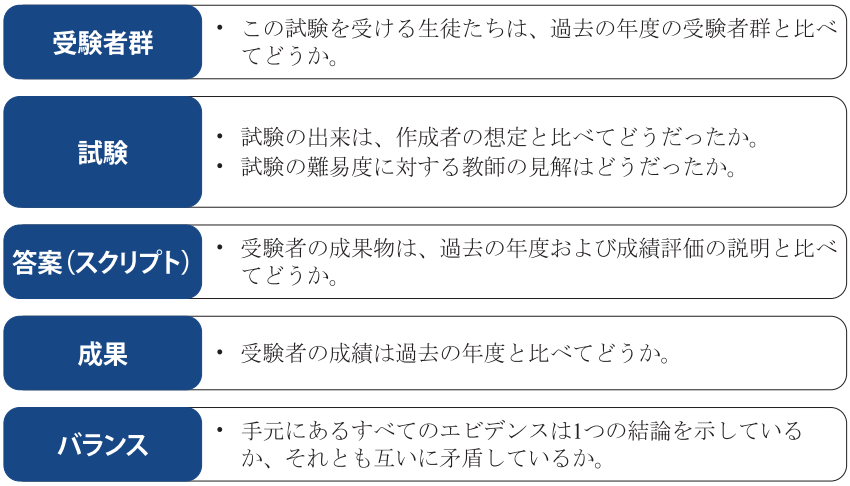
これは、主任試験官、試験官長、上級試験官、および IB が数日間にわたって話し合い、決定します。さまざまなエビデンス（詳細は「成績付与のプロセスで使われるエビデンス」のセクションを参照）を検討し、その年の生徒だけでなく、その前年に同じ科目で試験を受けた生徒にとっても公平な結論を導き出します。そして最後に試験官長が、この試験セッションの結果をどのようにすべきかについて IB に提言します。

この話し合いの主な成果として、成績区分、つまり各成績を付与されるために生徒が獲得しなければならない評点の最低ラインが決められます。すべての科目のすべての成績について詳細な判断を下すことは現実的ではないため、IB は試験官に対して「判断に基づく成績区分」をいくつか推奨するよう求め、残りの成績区分は計算で導き出します。詳細は、「判断に基づく成績区分と補間的な成績区分」のセクションを参照してください。

成績区分は各試験セッションで同一になるとは限りません。これは、生徒に出題される問題が試験ごとに異なるためです。成績区分は課題の難易度に応じて調整する必要があります。IB は、試験の難易度が毎年同じになるよう、あらゆる合理的な措置を講じていますが、IB の試験がもつ重要性を鑑み、試験問題が公開されてしまうというリスクを回避するため、試験問題を事前にテストするということは行っていません。

成績区分を決定するうえで、試験官と IB は図 32 に示す要素を検討します。

図 32
成績付与のプロセスで検討される要素



多くのケースで、IB は複数の評価要素を組み合わせて 1 つの全体的な成績を付与します。これは本資料の「集約」のセクションで詳しく説明されています。成績付与のプロセスにおいて試験官が最も留意しなければならないのは、何より重要なのは**全体的な結果**である、という点です。全体的な成績を公平にするために必要とされる場合は、個々の評価要素の成績区分の精度を落とすこともあり得ます。

成績付与のプロセスの正式な目的は以下のとおりです。

- 評価のために提出された成果物の分布において、成績評価の説明がその質を最も的確に表している範囲が変わる箇所を特定する
- 各評価要素について成績区分となる評点を決定する
- それらの成績区分の組み合わせが、科目全体で見た場合に公平な成績付与となることを確認する

以下の結果が得られた場合、成績付与のプロセスが成功したと言えます。

- 受験者群の情報を考慮に入れ、答案に関する判断と結果を示すエビデンスについて、おおむね合意が得られている
- 学校のパフォーマンスが大きくばらついている場合、その理由を説明できる
- 試験官長と IB の最高評価責任者が、評価の基準が維持されたと確信をもって言える

判断に基づく成績区分と補間的な成績区分

判断に基づく成績区分は、成績付与プロセスにおける話し合いを基に試験官長が提言するものです。MYP、DP、CP について、この判断に基づく成績区分は、それぞれ 2 と 3 の区分、3 と 4 の区分、6 と 7 の区分に対応します。残りの区分（1 と 2、4 と 5、5 と 6）については、判断に基づく成績区分に従って、計算によって導き出されます。これは、**補間的な成績区分**として知られています。

成績 5 または成績 1 を付与される生徒の割合に大きな変化が見られた場合、受験者群という文脈において議論される必要があります。その結果、判断に基づく成績区分の再検討、または、まれなケースでは、この境界線上にある評価用成果物の見直しという措置がとられます。

成績付与のプロセスで使われるエビデンス

成績の付与はエビデンスに基づくプロセスであり、以下を含むさまざまな情報が検討されます。

- 評価に対する教師からのフィードバック
- 生徒の成果物のサンプルに対する試験官の専門的な判断
- 今年の生徒の結果と過去の生徒の結果を比較する統計情報の確認

特定の種類のエビデンスが、他のエビデンスよりも重要だということはありません。すべてのエビデンスをバランスよく考慮して、結論を導き出す必要があります。

今年の受験者群を考慮する

成績付与プロセスの最初の作業は、評価を受ける生徒の集団が、過去の受験者群とどの程度類似しているかを考えることです。今年評価を受ける生徒が在籍する学校が、過去とおおむね同一であれば、パフォーマンスの差は試験の難易度によるものだと考えられます。これにあてはまらない場合（例えば、初めてその科目の試験を受ける学校が多数含まれている場合や、再受験者の割合が高い場合）、IB は、パフォーマンスの差を成績に反映させるかを判断します。

受験者群を比較する際に検討すべき要因の例として、以下が挙げられます。

- 評価を受ける生徒数の変化
- 英語、フランス語、スペイン語、またはその他の言語で受験する生徒の割合の変化
- 新しい学校数、およびその学校の生徒数
- 生徒が選ぶ選択項目における変化

評価についてのフィードバック

次の作業は、評価のパフォーマンスが想定と比べてどうだったかを検討することです。特定の設問が想定よりも大幅に難しかった場合や、成績 6 と成績 7 の区別を図ることが想定されていた設問でそれが実現されなかった場合、成績付与チームは、この点を考慮に入れて成績区分を決定する必要があります。

この話し合いのベースとなるのが、試験についての教師からのフィードバック、および個別の設問や項目に関するパフォーマンスの統計です（図 33 を参照）。チームには各試験官が作成した要約レポートも提供され、それに加えて、当該セッションにおいて生徒の成果物を採点した自らの経験を参考にすることもできます。

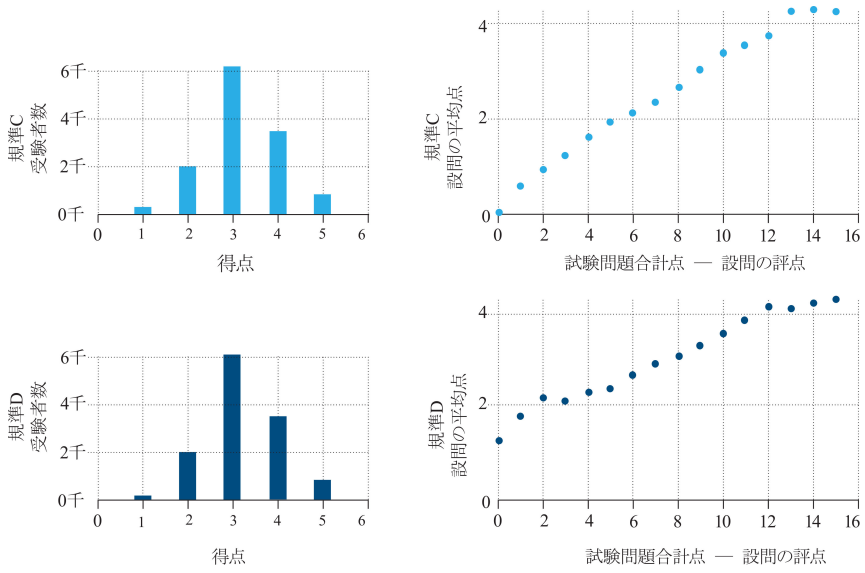
図 33
試験官に提供される項目別の統計の例

項目別の統計要約

設問	解答を試みた受験者数	受験者総数	解答を試みた受験者の割合	設問項目の平均点	項目の最高得点	設問項目評点の母集団標準偏差
01 規準A	12,071	42,042	28.71%	2.7	5	0.95
01 規準B	12,071	42,042	28.71%	2.6	5	0.94
01 規準C	12,071	42,042	28.71%	2.8	5	0.84
01 規準D	12,071	42,042	28.71%	3.2	5	0.79
02 規準A	29,848	42,042	71.00%	2.8	5	0.98
02 規準B	29,848	42,042	71.00%	2.7	5	0.93
02 規準C	29,848	42,042	71.00%	2.9	5	0.83
02 規準D	29,848	42,042	71.00%	3.2	5	0.79

項目と評価要素の残りの部分との相関

設問	相関係数
01 規準A	0.8044
01 規準B	0.8018
01 規準C	0.7983
01 規準D	0.7253
02 規準A	0.7954
02 規準B	0.7944
02 規準C	0.7865
02 規準D	0.7320



答案のエビデンスを確認する

次に上級試験官が、それぞれの成績評価の説明と、予想成績区分付近の評点を与えられた答案を比較します。成績付与プロセスのこの段階では、付与された評点ではなく、生徒の答案の性質、およびそれが成績評価の説明とどの程度一致しているかに焦点をあてることが大切です。

答案レビューを始める前に、試験官が過去の成果物の例に目を通し、それぞれの成績で期待されていることを改めて確認することは有益です。成績付与プロセスにおいて、このような答案が用意されている必要があります。

成績付与会議に先立ち、上級試験官チーム（特に主任試験官）が、各評価要素の暫定的な成績区分を提出します。これは、期待される基準に関するこれまでの経験から、成績区分を設けるべき評点についての試験官たちの考えを表すものです。この暫定的な成績区分とあわせて、各試験がどのように機能したかに関する合意、そして評点の全体的な分布に関する認識が提供されることで、成績付与のサンプルに表される評点範囲を提案するために必要な情報とエビデンスが、IB のサブジェクトマネージャーに与えられます。

各上級試験官は、所定の生徒の答案を確認し、その成果物の質を最もよく表す成績評価の説明を決定する必要があります。これは簡単な作業ではありません。たとえ解答の質が

比較的一定の生徒であっても、解答全体を見た場合、その質が複数の成績に幅広くまたがることが多いためです。ある答案が成績評価の説明にちょうどあてはまるのか、それとも上の成績に届きそうなのか（例えば、「7-」「6+」などで表すことが多い）を示すことは認められており、多くの場合で有益な情報となります。

この作業に取り組む試験官は、思い込みやバイアスをできる限り排除する必要があります。この理由から、各試験官が各自の結果を記録し終わるまでは、互いの見解について話し合わないようにすることが大切です。同様に、この段階が終わるまでは、統計分析の結果を試験官に知らせないようにすべきです（試験官長は除く）。ただし、個別項目における受験者群全体のパフォーマンスを示す関連性の高い情報については、この限りではありません。

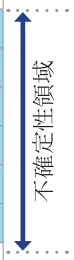
ある答案に最もあてはまる成績を決定することは非常に主観的な作業であり、試験官によってその結果がばらつく可能性があります。さらに、深い理解を示しているものの間違いが多い生徒と、簡単な問題では高いパフォーマンスを示しながらも理解の深さが劣る生徒がいる場合、後者の生徒の方が評点はわずかに高いにもかかわらず、前者の生徒の方に高い成績が付与されるということも起こりえます。また、答案レビューにおいては、比較的少数の生徒の成果物しか検討しないということも重要な点です。

試験官は、自分が見ている他の答案と質の異なる答案を特定することに最も長けているという結果が研究によって示されています（本資料の「採点に対するアプローチ」のセクションを参照）IB は、サンプルセットのうち、評点が最も高いものから作業を開始し、高い成績を示すエビデンスが一貫して見られなくなったと感じるまで、評点が低い方へと段階的にレビューを続けるように指示しています。その後、同じサンプルセットの中で評点の最も低いものをレビューし、高い成績を示すエビデンスが一貫して見られるようになるまで、その作業を続けます。必要に応じて、サンプルセットの評点の範囲を広げることができます。

すべての上級試験官の成績を集めると、成績区分を設定すべき評点の範囲が浮かび上がります。これは「不確定性領域」と呼ばれます。成績区分を設定するだけの合理的なエビデンスが存在する評点の最低レベルと最高レベルを表しています。不確定性領域を設定する方法については、明確に決められていません。これは、上級試験官たちがそれぞれの判断を基に、話し合いを通して合意すべきものです。

図 34
答案レビューの結果の例

答案番号	評点	試験官1（主任試験官）	試験官2	試験官3	試験官4
1	51	4	4	4	4
2	51	4	4	4	4
3	50	3+	4	4	3+
4	50	4-	4	4	3+
5	49	4-	4-	4-	4-
6	49	3+	3	4-	3+
7	48	4-	3+	4-	3+
8	48	3+	4-	4-	3+
9	47	3	3	3+	3
10	47	3	3	3	4-



このプロセスの唯一の例外は、多肢選択式の試験問題です。これまでの経験から、評価のために提出された成果物の質に基づいて成績区分を判断することは、多肢選択問題から構成される試験では非常に難しくなります。これは、実際に生徒が達成したことを判断するためのエビデンスが、解答にほとんど含まれていないためと考えられます。このような試験では、各成績内の生徒の割合が、最も関連性が高いと考えられる試験において判断に基づいて確立された割合にできる限りあてはまるように、計算によって成績区分が求められます。

結果の統計を確認する

専門家の判断に依存する規準準拠型の評価では、上級試験官が、試験の設問および各設問が生徒の解答に求める内容のみを考慮に入れて成績区分を設定することができるはずだと考える人もいます。ただし、実際には、生徒がどのように解答したかを見ることなくそのような判断を下すことには大きな困難が伴います。クレスウェル（Cresswell, 2000）は、成績を付与する立場の者は、試験セッション同士を比較して難易度が高い部分と低い部分を特定することに関しては正しい判断ができるものの、難易度がどの程度高かったのか、また低かったのかについては、往々にして正しく見積もることができないと結論づけています。

成績付与の目的は、成績の意味や基準の一貫性を保つことです。そのため、ここでクレスウェルが提唱した同等性の定義の 1 つを確認することには、大きな意義があります。

> 能力および過去の到達度の分布が同じであり、同一の入学方針をもつ類似の学校に通い、同等の能力をもつ教師から指導を受け、同等の意欲をもつ 2 つの生徒群が、それぞれのシラバスを学習し試験を受けた後に受け取った成績が同一の分布となった場合、2 つの試験の水準が同等であると定義できる可能性があります。
>
> （Cresswell, 1996, pp. 57-84）

成績付与のための参考情報として、上級試験官には**統計的推奨区分**が提供されます。これは、所定の成績を付与される生徒の累積率（所定の成績以上の成績を付与される生徒の割合の合計）が、前年と同じになるよう設定された成績区分と定義することができます。

成績付与のプロセスに関与する試験官には、生徒が獲得した評点の中央値および実際の評点分布を表すヒストグラムも提供されます。この情報は統計的推奨区分に加味されているものの、このような詳細情報を検討できるということは、多くの場面で有益となります。

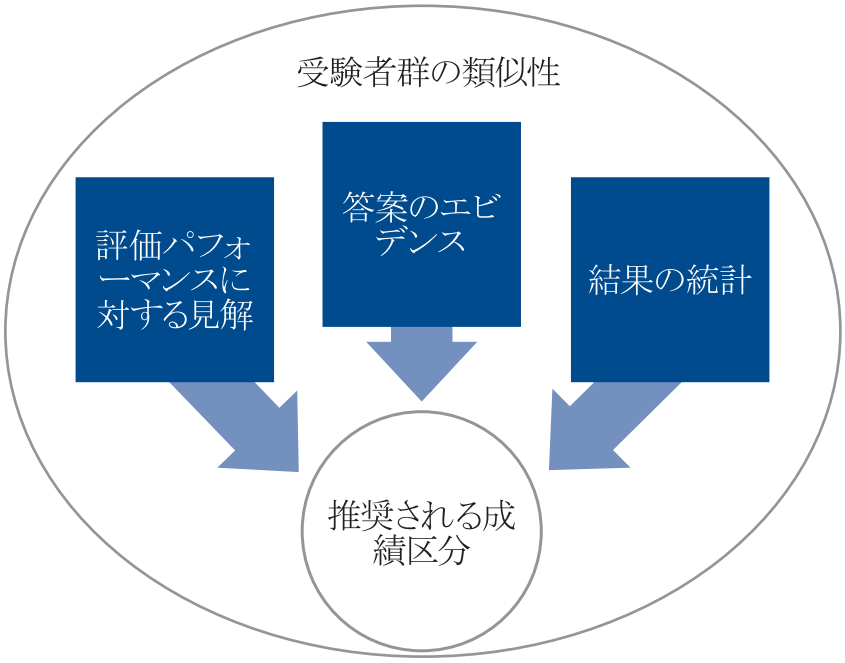
試験を受ける生徒群はセッションごとに異なります。したがって、上記のような同等性の定義の前提が完全に満たされる可能性は低くなります。これは特に、2 つの受験者群が互いに大きく異なる場合にあてはまります。同様に、同じような教育体験（例：IB ワールドスクール）をもつ多数の生徒に対応する場合、成績が大きくばらつく原因は、生徒のパフォーマンスの差ではなく、生徒が異なる設問群に解答したためだと考えられます。

統計的推奨区分の意義を検討する際には受験者群の規模を考慮する必要がありますが、これに関して IB では正式なルールを設けていません。他の要因も重要であるためです。一般的に、受験者が数十人の場合、全体的な成績に大きなばらつきが見られる可能性がありますが、数千人規模の受験者群ではばらつきは比較的少なくなります。

エビデンスのバランスをとる

特定の種類のエビデンスが他のエビデンスよりも重要だということはありません。そのため、成績付与チームが成績区分について提言を行う際は、すべてのエビデンスをバランスよく考慮する必要があります。

図 35
成績区分の選択を支えるエビデンス



複数の異なるエビデンスが同じ結果を示すというのは珍しいことではありません。例えば、統計的推奨区分が不確定性領域に含まれる場合などがこれにあたります。ただし、複数のエビデンスが矛盾することもあります。このような場合、最終的な判断の根拠を示す際に、この矛盾を説明することが非常に重要です。

これは成績付与プロセスの最終段階であるため、さまざまな判断の違いが、受験者群の全体的な結果にもたらす影響をシミュレーションすることが可能です。成績付与チームは、重要なのは個々の評価要素の成績ではなく全体の成績であることを踏まえ、比較の信頼性を理解するための過去のデータを参照しながら、受験者群の成績の中央値と教師の予測スコアとを比較することができます。また別のアプローチとして、新しい学校のパフォーマンスが経験の長い学校のパフォーマンスと大きく異なる場合、新しい学校を全体的な結果から除外するということもできます。

最後に、成績付与チームが提言およびその根拠をまとめて提出し、承認を受けます。その後、集約が実行されます。これは、さまざまな評価要素の評点と成績区分を合算し、最終的な評点または全体的な成績区分を確立するプロセスです。これを実現するため、全体的な評価要素の評点や成績区分を換算する必要が生じることもあります。詳細については、本資料の「集約」のセクションを参照してください。

固定された成績区分

大部分の内部評価課題および一部の外部評価課題においては、どのセッションでも実質的に同じ課題が生徒に与えられます。例として、「芸術」のポートフォリオや「パーソナルプロジェクト」などが挙げられます。このような場合、IB は（過去の成果物の標準化を通して）毎年同じ採点基準を維持するため、成績区分も変わらないと考えることができます。

成績付与プロセスにおいて、主任試験官と試験官長は、既存の成績区分が適切でないことを示すエビデンスが認められるかどうかを検討します。大抵の場合、既存の区分を持ち越すべきであるという結論が出されることになります。

ただしここで明確にすべきなのは、成績区分の検討は毎年実施されており、セッション間の成績の同等性を維持することができないというエビデンスが認められた場合には、調整が加えられるということです。このような状況は、生徒の行動に大きな変化があった場合（課題を完了するためのテクノロジーの新規開発など）や、提出された成果物が変化したことで課題の全体的な質が維持されたままマークスキームとの整合性が高まった場合、また、IB における成果物の採点やモデレーションの方法が見直され、その結果採点基準が変更された場合などに起こることがあります。

成績区分の変更の理由やその意味についての詳細は、「Setting grade boundaries: Why grade boundaries can change, and what it means if they do（成績区分の設定：成績区分はなぜ変更されるのか、その変更は何を意味するのか）」（動画）を参照してください。

推奨された成績区分の確認と承認

成績付与チームによって、成績区分をどのように設定すべきかについての提言がまとめられると、この提言内容及びその根拠が **24 時間配信科目レポート**として提示されます。このレポートには、コースを履修した生徒群における変化や、提案された区分に基づく成績結果を過去と比較したデータなど、裏づけとなる情報も含まれます。その後、このレポ

ートを IB 評価部門の上級職員が精査し、議論に十分な説得力があるかどうかを決定します。

IB 評価部門のリーダーシップチームが提言内容を受諾できない場合、サブジェクトマネージャーと懸念事項について話し合い、エビデンスの 1 つの側面をより集中的に掘り下げたり、提言内容を裏づける分析をさらに提供したりするよう求めます。

成績区分の設定に関する最終的な承認は、IB の最高評価責任者が、成績付与チームからの提言に基づいて行います。

プログラム認定証の授与

IB プログラムの認定証（MYP 修了証、IB ディプロマ、CP 修了証）の授与については、成績付与プロセスではなく、成績開示の準備プロセスにおいて決定されます。このプロセスについては、「成績の開示に向けた準備」のセクションで説明されています。

教師オブサーバー

IB は、評価プロセスの透明性を高め、成績が付与される方法に関する一般的な理解を深めることに取り組んでいます。この取り組みの一環として、教師オブザーバーを成績付与会議に招待しています（正確な詳細は、会議が対面かオンラインかによって異なります）。オブザーバーとなった教師には、学校の同僚教師にその体験を報告するとともに、IB 教育者ネットワーク（IBEN：IB educator network）の広いコミュニティーにレポートを提供することが期待されています。詳しくは、IB アンサーに問い合わせてください。

成績付与プロセスの原則

IB の成績付与プロセスの基本となる原則は以下のとおりです。

- 利用できるすべての（判断的および統計的）エビデンスを使って、3 と 4、6 と 7、2 と 3 の成績区分をこの順番で決定する。提出された成果物がなく、成績区分が設定できない場合は、利用できるエビデンスを基にどの区分が最も適切かを決定する。
- 他の成績区分については、適切な手順に従って計算で求める。
- 成績区分の決定は、試験官の判断によるエビデンス、統計的なエビデンス、受験者群の情報を使った三角法の原理を基にする。3 つすべてが均等なバランスになるよう、妥協点を見つけながら調整する。
- ある評価の受験者群が過去の受験者群と全体的に類似している場合、IB はその結果も過去の受験者群と全体的に類似すると考える。ただし、以下を考慮する必要がある。
  - 多くの場合、受験者群は毎年異なる。これは履修者が少ない科目に特にあてはまる。
  - 結果に差がある場合、IB は、その根拠として確固たるエビデンスが存在するものと考える。
  - ある評価の課題が過去のものと全体的に類似する場合、IB はその成績区分も過去のものに類似すると考える。

- すべての成績区分は（内部評価の課題であっても）毎年変わる可能性がある。
- IB は、評価の難易度が毎年一定になるようあらゆる措置を講じているものの、特定の試験の要求度が変化することを認識している。
- 全体的なコースの成績区分が優先される。生徒にとって重要な意味をもち、関係者の意思決定に使われるのは全体的な成績である。
- 評価要素の成績区分は、全体的なコースの成績を堅牢なものとするための重要なステップである。ただし、評価要素レベルでの小さな影響が組み合わさり、全体的なコースの成績に大きな影響を与えることがある。

品質確認

- 評価の最も重要な成果は生徒が受け取る成績です。そのため、最終確認では、この成績に重点が置かれます。
- 最終確認の目的は、異常を見つけることです。

成績区分が設定された後、生徒に付与される成績が公平かつ適正であることを担保するため、追加の品質確認が行われます。

この確認の 1 つに、「アットリスク（要確認）」と判定された生徒の再採点があります。これは、最終成績が想定よりも低かった生徒を特定するものです。「アットリスク」の再採点の一部では、その採点を詳しく調査する必要があるとされた試験官の作業が集中的に見直されます。再採点は、主任試験官の採点基準を一貫して適用できることが証明された試験官によって実施されます。「アットリスク」の再採点を行う目的は、加点できるかどうかを判断することではなく、現在の評点が適切かどうかを確認して、生徒が正しい成績を確実に受け取れるようにすることです。

IB の評価スタッフもチェックを行い、学校の科目の成績が適切と見なされることを確認します。

IB の資格授与委員会

- IB の資格授与委員会は、MYP、DP、CP に関連するさまざまな関係者を招いて、評価セッションについての概要を説明し、その精査を行う公式な場です。PYP には、確認や認定が必要な最終成績が存在しないため、資格授与委員会もありません。
- IB の資格授与委員会は、組織全体から集まるさまざまな代表者から構成され、ここには IB の上級職員や試験官長も含まれます。成績開示の最終的な承認を行うとともに、IB の学問的誠実性に関する方針に違反するケースや、その他の関連する問題が検討されます。

IB の資格授与委員会は、IB に提出された評価による、IB プログラムに関する生徒の成績承認の最終段階です。この委員会において、セッションが IB の基準を満たしていたかどうか、および IB 理事会に代わって成績を付与すべきかどうかについて、最高教育責任者や評価主任の提言が確認されます。また、IB 資格の授与に関する方針や前例を設定する場でもあります。

委員会の規模や構成はプログラムによって多少異なりますが、投票権をもつ委員の数は同じです。委員として参画するのは以下の人物です。

- 教育局および学校局を担当する IB の上級職員
- さまざまな科目の試験官長

成績区分の承認の責任を負うのは評価担当の上級職員ですが、IB の資格授与委員会は、その監督機関として機能し、全体的なプログラムの修了率といったマクロレベルの結果を確認するとともに、IB の評価担当から報告されたあらゆる問題を検討します。

また、学問的不正行為および学校の過失に関する問題の確認も行います。これについては、学問的誠実性を担当する小委員会で詳細に議論されます。

IB の資格授与委員会の最後の役割は、試験セッションのパフォーマンスを振り返り、次回以降のセッションに向けた IB に対する提言をまとめることです。

利益相反

IB の資格授与委員会は重要な意思決定機関であり、透明性と独立性を維持することが重要です。特定の議題について利益相反となる可能性がある委員は、その議論に加わらず部屋を出る必要があります。このことは、すべての会議の冒頭で必ず参加者に伝えられます。

オブサーバー

IB では、資格授与委員会へのオブサーバーの参加を奨励しています。その目的は以下のとおりです。

- 一連の手順の透明性を高める
- プロセスの変更・改善点について提案を行う機会をつくる
- IB の関係者同士のパートナーシップを認識する

IB が上記の目標を達成できるよう、オブサーバーは会議に参加してから 2 週間以内に、最高教育責任者に対して報告書を提出するよう要請されます。この報告書には、現在の手順に関する全体的な所見、および変更・改善すべき点がある場合は、それについての提案などが含まれます。委員会が検討した個別の案件についてのコメントは差し控えるべきです。

委員会で取り上げられる問題は機密情報であるため、オブザーバーには守秘義務および個人的な利益相反に関して、適切な制約が課せられることになります。オブサーバーには、委員会での決議に関して投票を行う権利はありません。

IB の資格授与委員会にオブサーバーとして参加したい場合は、support@ibo.org に問い合わせてください。

成績の開示に向けた準備

- すべての評価が採点され成績区分が決定された後、生徒の成績を開示する準備を整えるために、いくつかのプロセスを完了する必要があります。
- 成果物が評価のために提出されたというエビデンスが存在しない場合、IB は「評点がつけられない場合の手順」（missing mark procedure）を使って評点を計算します。
- 各プログラムで資格取得を目指している生徒については、MYP 修了証、CP 修了証、IB ディプロマを授与すべきかどうかを決定するために、科目の成績が合算されます。
- 成績の公開には、堅牢な変化点管理システムが用いられます。成績に加えられた変更はすべて把握され、該当する関係者（学校または大学）に情報が共有されます。

成績区分の設定と生徒への成績の開示の間には、完了すべきプロセスや手順がいくつもあります。この手順やプロセスには、単年度履修科目の持ち越しといった一部の生徒にしか影響しないものもあれば、すべての生徒に適用されるものもあります。

評点がつけられない場合

- 学校や生徒の過失以外の理由で、IB が生徒の成果物を確認できない場合、生徒が不利益を被る可能性を最小限に抑えるため、評点を推定します。
- この推定はエビデンスに基づいて行う必要があります。評価のために提出された成果物の数が非常に少なく、推定ができない場合は、別の方法で評点を決定する必要があります。
- 評点がつけられない場合の手順は、生徒の平均的なパフォーマンスに基づきます。そのため、この推定によって利益を受ける生徒の数と、不利益を被る生徒の数は、おおむね同数となります。どのような場合でも、最も公平なのは、評価のために提出された実際の成果物を採点することです。

学校や生徒がコントロールできない何らかの理由により、IB が生徒の成果物を採点できなくなるという状況が時おり発生します。例えば、紙ベースの試験が郵送途中で紛失する、または、試験当日に生徒が突然体調を崩すといった状況が考えられます。

評点がつけられない場合の手順は、このような状況において、生徒の評点を推定するために使う仕組みです。

IB または（学校以外の）第三者の行動の結果として生徒の成果物が利用できなくなってしまい、もう一度評価を受けるよう要請することが合理的でない場合には、この手順をとることが適切です。

評点がつけられない場合の手順はすべて、生徒の到達度を示すエビデンスに基づいて行う必要があります。生徒のパフォーマンスに関する情報が不足している場合、公平な推定評点を導き出すことが難しくなります。

評点がつけられない場合の手順では、所定の評価要素におけるすべての生徒の到達度の平均を、他の評価要素における生徒のパフォーマンスと比較した結果を基準にします。この平均をとることで、実際よりも低い評点となる生徒の数と、実際よりも高い評点となる生徒の数が同じになると考えられます。そのため、いかなる場合でも、実際の生徒の成果物を採点することが最も公平な方法です。

同じ科目とレベルに登録した生徒が 5 人以上いる場合は、学校のデータを使って評点を決定します。校内の登録生徒数が 5 人に満たない場合は、グローバルデータを使って評点を見積もります。

図 36
評点がつけられない場合に評点を決定する

評点がつけられない場合に受験者の評点を決定する方法:

$$\text{比率} = \frac{\text{当該受験者の他の評価要素の評点換算点の合計}}{\text{他の評価要素のグローバル (全校) 評点換算点の合計}}$$

$$\text{受験者の評点} = \text{評点がつけられない評価要素のグローバル (全校) 評点換算点} \times \text{比率}$$

この比率は、生徒が完了した評価要素について、生徒の評点と世界平均または学校平均とを比較するものです。

- 生徒の評点が平均を上回る場合、比率は 1 よりも大きくなります。
- 生徒の評点が平均を下回る場合、比率は 1 よりも小さくなります。

この計算は、評点がつけられない場合の他のすべての手順と同じく、利用できるエビデンスに基づいた「最も信頼できる予測」です。どのような場合でも、一番望ましいのは生徒の実際の成果物を採点することです。

成績がつけられない場合の手順

MYP では、特定の科目の評価モデルに評価要素が 1 つしかない場合、評点の推定に使えるエビデンスが存在しないため、評点がつけられない場合の手順を使うことはできません。この状況では、別のアプローチとして、MYP での成績がつけられない場合の手順を使います。

評価要素を 1 つに絞ることで生徒の作業負担を管理することができますが、MYP では、利用できるエビデンスが制限されてしまうというデメリットが生まれます。これは、成績がつけられない場合の手順に対する IB の信頼度が、他のプログラムにおいて評点がつけられない場合の手順に対する信頼度に比べて低くなることを意味します。したがって、本手順は例外的な場合にのみ使用すべきです。

生徒が要件に従って評価を完了し、それを正当な意図をもって提出したものの、生徒や学校がコントロールできない何らかの要因によってその成果物を IB が採点できない場合、IB は成績がつけられない場合の手順を適用します。

成績がつけられない場合の手順を適用するためには、生徒が MYP において少なくとも 4 つのコースで最終的な成績を付与されている必要があります。それよりも少ない成績しか付与されていない生徒については、情報に基づいた推定をするためのエビデンスが足りないため、成績を見積もることができません。

成績がつけられない場合の計算では、過去 18 か月に評価のために提出された成果物を基に決定された成績をもって、他のすべての科目の成績の中央値を求めます。

成績の中央値が 0.5 以上の場合は次の整数に繰り上げます。

成績の中央値が 0.5 未満の場合は繰り下げます。

この計算は、評点がつけられない場合の他のすべての手順と同じく、利用できるエビデンスに基づいた「最も信頼できる予測」です。どのような場合でも、一番望ましいのは生徒の実際の成果物を採点することです。

プログラムの成果

- IB は、個別の判断ではなく、規準を満たしているかどうかに基づいてディプロマや修了証を授与します。
- ディプロマや修了証の授与については、プログラムごとに独自の規準が定められています。

生徒の全科目の成績が出そろうと、IB は、その生徒が MYP 修了証、IB ディプロマ、または CP 修了証を取得する資格があるかどうかを計算することができます。

MYP、DP、CP の成果の計算方法について、および合格規準についての詳細は、IB ウェブサイトおよび各プログラムの『評価の手順』に記載されています。

成績の開示

IB の立場からすると、成績の開示において最も大事なことは、非常に厳格な変化点管理プロトコルを実施することです。

試験セッションにおいては、IB が常に状況を明確に把握できるようにするため、情報が継続的に更新されます。生徒と学校に成績が公開された後は、関係者全員に適切に通知することなく成績が変更されることはありません。DP と CP では、これは特に、生徒の成績証明書を受け取る大学にあてはまります。

成績開示後に成績の変更が承認されることがありますが、これは、成績照会サービスが利用される、保留中の成績が学校によって確認される、学問的誠実性の事例が解決する、といった理由によるものです。

成績照会サービスと評価に対する不服申し立て

- 成績照会サービスの目的は、採点において手違いがあったと学校が考える場合に、その旨を IB に通知できるようにすることです。
- 成績照会においては、元の試験セッションと同じ基準が適用されなければならず、IB はそれを確認するためにシード品質モデルを採用します。
- 外部評価は、形成的評価ではなく総括的評価として機能するよう意図されているため、試験官は、採点の助けとなる場合にのみコメントを残すように求められます。
- IB はまた、適切な措置が講じられなかったと考える学校が、不服申し立てを行うための正式なプロセスを設けています。

提供される成績照会サービスの法的および手続き的な説明については、各プログラムの『評価の手順』を参照してください。

成績照会サービスのカテゴリー

成績が開示された後、コーディネーターは以下のいずれかを申請できます。

- **カテゴリー 1 の成績照会** — 1 人の生徒の 1 つの科目におけるすべての外部評価要素の再採点
- **カテゴリー 1 のレポート** — カテゴリー 1 の成績照会により行われた採点結果の報告
- **カテゴリー 2 の成績照会** — 外部評価要素の評価済み成果物のコピー
- **カテゴリー 3 の成績照会** — 内部評価要素の再モデレーション

上記の各カテゴリーに対して料金が発生します（カテゴリー 1 再採点において成績が変更された場合を除く）。

成績照会サービスを提供する理由

成績照会サービスは、採点システムの間違いに対する最後の修正措置として設けられています。採点に間違いがあると考える学校がその旨を IB に伝え、IB がその件を調査のうえ、必要に応じて間違いを訂正する機会を提供するものです。

ライブマーキングと成績照会サービスの採点の間で基準を維持する

成績照会サービスの目的は、IB の基準に基づいて、成果物の質を反映する評点を生徒に与えることです。最初に付与された成績に対して生徒が不満をもっているという事実、または生徒の成績が 2 つの成績区分の境界付近にあるという事実が、試験官に不適切に影響しないようにすることが非常に大切です。

成績照会サービスのプロセスでは、最も一貫性をもって採点できる上級試験官のみが採点にあたります。そのため IB は、この再採点の結果が正しい評点であるとみなします。

試験官が基準を維持できるよう、IB はシードスクリプトを成績照会サービスの成果物の中に含め、試験官が合意済みの基準から逸脱し始めた場合に、電子的にその旨を通知します。

IB の評価スタッフは、成績照会サービスで提案された成績の変更をすべて確認し、新しい成績が元の成績よりも信頼性の高い結果であることを確かめます。2 人の試験官の評点が明確な理由なく大きく異なり、それによって成績が変更される場合、第三者による答案の確認が要請されることがあります。

カテゴリー 1 の成績照会サービス：再採点

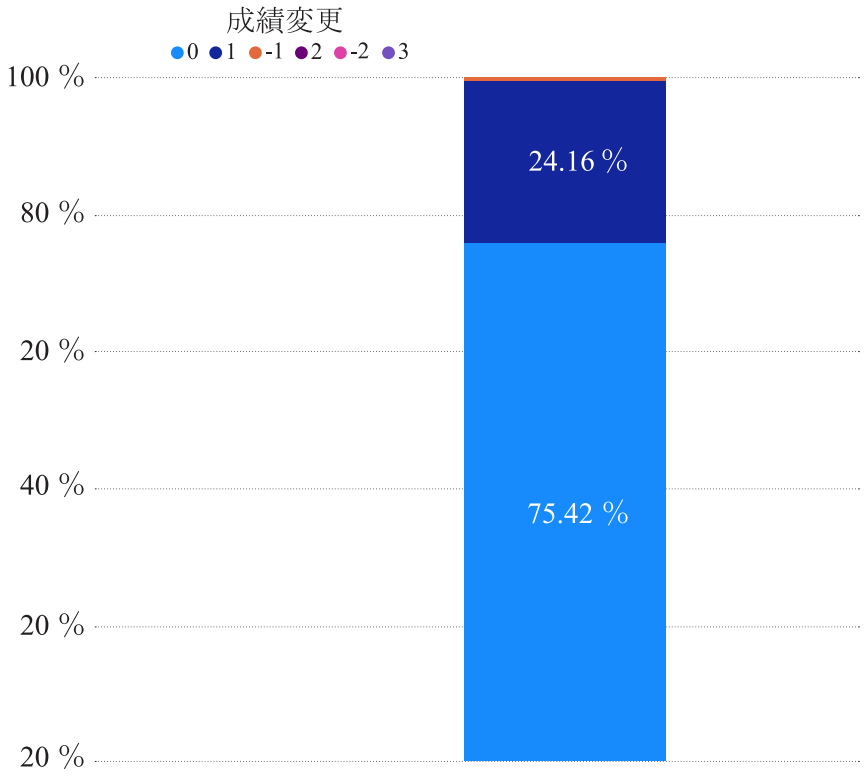
「採点」のセクションで説明したように、2 人の試験官が、許容される範囲内で同じ設問に対して異なる評点をつけることがあります（許容差）。そのため、成績照会サービスの結果として、評点がほんのわずかしか変更されなかった場合、採点には間違いがなかった可能性が高くなります。IB の成績照会サービスの申請では、大多数がこの結果となります。

試験官の評点におけるこのわずかな（かつ許容される）差異の影響を最小限にするため、カテゴリー 1 の成績照会サービスでは、1 人の生徒の外部評価要素がすべて再採点されます。評点の違いが体系的な問題によるものではない場合、複数の試験を合わせるとその差異がプラスマイナスゼロとなると考えられるため、小さな評点変更の影響を抑えることができます。

IB は、成績照会サービスの結果として出される評点が、生徒の成果物を最も正確に反映したものだと確信しているため、これが生徒に付与される最終的な評点となります。これは、カテゴリー 1 の成績照会サービスの結果、生徒の成績は上がるだけではなく下がる可能性があることを意味します。したがって学校は、カテゴリー 1 の成績照会サービスを申請する前に、生徒の同意を得る必要があります。

成績区分の設定プロセスの性質により、付与された評点が区分境界の 1 点下、または 1 点上となる生徒が必ず出てきます。このようなケースでは、採点には誤りがなかったにもかかわらず、成績照会サービスの結果として成績が変更になることがあります。この際 IB は、その生徒の評点が 2 つの成績の境界にあること、そして、その 2 つのうちのどちらがそのパフォーマンスを公正に表しているかを明確に伝えます。

図 37
カテゴリー 1 の成績照会サービスによる成績変更（DP の 5 つの試験セッションの平均）



カテゴリー 2 の成績照会サービス：生徒の成果物の返却

カテゴリー 2 の成績照会サービスの対象となる評価は、総括的評価です。返却された答案には、担当試験官からの有益なコメントが含まれている可能性があります。ただし、試験官が生徒の成果物を採点する際にコメントを残すことは義務ではないため、必ずコメントが提供されるとは限りません。

IB は、提出された成果物を反映する公正な成績を生徒に付与するため、適切な採点を行うことを試験官に求めています。採点の役に立つコメントは記入するよう指示していますが、改善のためのフィードバックを生徒に提供することは必須ではありません。

成果物の採点が公正ではないと考える場合の措置については、「Challenging results: What you can do if you think your work hasn’t been marked fairly（再採点を求める：成果物の採点が公正ではないと思われる場合の措置）」（動画）を参照してください。

カテゴリー 3 の成績照会サービス：再モデレーション

カテゴリー 3 の成績照会サービス（再モデレーション）では、評点の変更はすべて受験者全体に適用されます。IB が、生徒の同意を得るという原則を適用した場合、1 人の生徒が同意を拒否する、または同意依頼に返答しないために、受験者群全体の成果物の再モデレーションを行うことができないという事態が発生することになります。これは明らかに道理にかなっているとは言えません。そのため、成績が下がるリスクを正式に受諾してい

ない生徒の成績が下がってしまうことがないよう、カテゴリー 3 の成績照会サービスについては、成績は上がるのみで、下がることはありません。

再モデレーションは、生徒の内部評価のモデレーション後の評点平均と、元の評点（教師がつけた評点）の平均との差が、その評価要素の最高評点の 15%以上となった場合にのみ行われます。

このようなしきい値を設けているのは、仮に IB が一切制約を設けなかった場合、成績が上がることを期待して、全科目について再モデレーションを常に求めてくる学校が出てきてしまうからです。IB の上級モデレーターの数は限られているため、このような要請に対応することはできません。このようなしきい値を設けることで、明確な懸念をもつ学校のみが再モデレーションサービスを利用できるようにしています。これにより、少数しかいない上級モデレーターの業務量を管理しつつ、生徒の評点に対するモデレーションの影響が大きかった学校に対しては再モデレーションの機会を提供することが可能になります。

図 38
カテゴリー 3 の成績照会サービスを申請することができる受験者群

生徒番号	サンプル対象	モデレーション前の評点	モデレーション後の評点	モデレーションによる調整
2	対象	19	14	-5
10		15	12	-3
11	対象	16	13	-3
12		20	15	-5
13		18	14	-4
16	対象	12	11	-1
36	対象	17	13	-4
45	対象	10	10	0
	評点の平均変化			-3.125
	最高評点に対する平均変化の割合			-15.63 %

図 38 の例では、この受験者群に付与できる最高点は 20 点です。

モデレーションによる評点調整の平均値は-3.125 点であり、最高点の 15.63%に値します。

つまりこの例では、カテゴリー 3 の再モデレーションを利用することができます。

評価に関する不服申し立て

評価に対する不服申し立ては、生徒の IB の成績の処理や生徒による学問的不正行為の処理など、IB の評価プロセスにおける意思決定過程の再調査を依頼するための手段です。

不服申し立てにより IB のプロセスに例外を適用するよう求めることはできませんが、IB の規則の解釈の合理性に疑問を呈することは可能です。不服申し立てにより、発行済みの IB の方針を変更することはできません。不服申し立ての結果、IB の規則が明確化された

場合は、明確化された解釈を同セッションのすべての生徒に適用する必要があります。不服申し立ては、IB の評価プロセスに関する最終的な措置です。

詳細は、プログラム・リソース・センターで入手できる MYP、DP、CP の『評価の手順』を参照してください。

再試験生

生徒は、付与された成績が満足のいくものでなかった場合、その科目を再度受験することができます。これは、6 か月後または、原則的にはそれ以降のいかなる試験セッションでも行うことができます。ただし、ある科目を再履修する際にカリキュラムまたは評価要件が大幅に改訂されていた場合、生徒はその新しい要件に従わなければなりません。1 つ以上の科目の再履修を希望する生徒は、最初にその科目を履修した学校と同じ学校に登録する必要はありません。学校が 1 つ以上の科目を再履修する生徒を受け入れる場合、その生徒に対するすべての学問的および管理上の責任をその学校が引き受けなければなりません。

科目を再履修する生徒は、カリキュラムや評価要件に大きな変更がなければ、試験以外の評価要素の結果を持ち越すことができます。ただし、筆記試験については採点結果を持ち越すことができないため、再履修する科目のすべての試験を再度受験する必要があります。

生徒が、試験以外の評価要素のために成果物の再提出を希望する場合、再試験セッションに登録した学校で授業を受けなければなりません。その理由は、科目担当教師が学問的な指導を提供し、内部評価用の成果物を採点し、すべての成果物が生徒本人が取り組んだものであることを確認しなければならないからです。以前に提出した成果物にわずかな変更を加えただけでは不十分です。評価のために提出する成果物は、原則として、完全に異なるものであるべきです。ただし、非常に規模の大きなコースワークの場合、時間的制約を考慮すると、これは必ずしも現実的ではないということを IB は認識しています。その場合、元の成果物に大幅な修正を加えたものも認められます。修正した成果物を提出する場合、その成果物は新たに採点されます。つまり、成果物の評点が下がる可能性もあるということを生徒は理解しなければなりません。修正した成果物には、元の成果物に費やした総時間数の 20%以上を追加で費やす必要があります。また、成果物が次の規準の少なくとも 1 つを満たさなければなりません。満たすべき条件：

- 研究課題（リサーチクエスチョン）、焦点、または命題を変更した
- 結論に影響する新しいテーマ、エビデンスのストランド、または視点を追加した
- 分析や計算を大幅に追加した
- 著者または作者の作品の側面について新しい（追加の）探究を行った

生徒が完成させたものの評価の一部（例：DP の「芸術」のポートフォリオ）として提出していない成果物がある場合は、この未提出の成果物を作成するためにすでに費やした時間を 20%の目標時間数に数えることは認められます。

このアプローチは、再試験生は成果物に対する追加のフィードバックから利益を得るべきではない（新しい成果物を提出した生徒が不利になるべきではない）という考え方と、

生徒や学校に過度な負担（再試験を確実に避けるために、まったく新しいデータを取得したり研究を行ったりするなど）を強いることはフェアではないという考え方のバランスをとるために採用されています。

同等性に関する IB のアプローチは、本資料の「同等性」のセクションに記載されています。

IB は、時間の経過に伴い、コースの内容を更新していく必要性を認識しています。新しいコースを導入する際は、プログラムの更新情報や科目の『指導の手引き』の改訂版の作成などを通して、さまざまな方法で変更点を事前に学校に伝えています。変更前のコースの最後の試験セッションが終了した後は、変更前のコースに関する試験を受け直す機会を再試験生に提供することはできなくなります。これは、同じ試験セッションの同じ科目内で 2 つのコースに関する試験を実施することは、学校にとっても IB にとっても、管理上の負担が大きすぎるためです。ただし、再試験生は、内部評価または外部評価のコースワークの評点を持ち越せる可能性があります。新しいコースが以前のコースと大幅に異なる場合、持ち越しが認められることはほとんどありません。

再試験生についての詳細は、各プログラムの『評価の手順』を参照してください。

学校へのフィードバック

- IB の総括的評価の目的は生徒のパフォーマンスを測定することであり、関連するすべてのプロセスは、この測定結果の妥当性を最大化するために設計されています。
- IB が評価結果を学校および生徒に伝える際には、透明性を重視し、評点と成績がバイアスのない一貫した方法で付与されていることを教師や生徒が確認できるようにしています。
- 教師の評点のモデレーションに基づいて成績が決まる内部評価では、透明性が特に重要になります。
- 成績照会サービスのプロセスを含む IB の総括的評価は、生徒の成績を向上する方法について学校にガイダンスを提供することを意図していません（ただし、妥当な評価を行うために必要なデータの副産物としてガイダンスが提供されることがあります）。
- IB は、評価プロセス以外のサービスとして、指導方法に関するサポートや教職員研修などを提供しています。

IB の総括的評価の目的は、学習プログラムの最後に生徒のパフォーマンスを評価し、信頼できる測定結果を成績という形で提供することです。フィードバックを通して生徒の学習を支えることを目的とする形成的評価とは異なります。意味のあるフィードバックは、形成的評価および効果的な指導に欠かせない要素ですが、総括的評価の結果を形成的な目的で使用することは適切ではありません。この理由の 1 つとして、IB の試験官は目の前にある 1 つの成果物、つまり試験の答案について判断を下すことになる一方、実際に教室で指導する IB の教師は、さまざまな情報と個人的な経験を参考にして生徒にフィードバックを提供できるということが挙げられます。そのため、試験官が採点時に生徒や教師に向けたフィードバックを記入することはありません。

試験官には、評点が付与された部分を明確に示し、曖昧さが残る場合には、その説明のために適切なコメントを追加することが求められています。これは、基準が守られていることを確認し、評点が付与された箇所について学校に透明性を提供するという点で、IB をサポートするものです。

科目レポート

各試験セッションが終わると、IB は科目レポートを発行します。このレポートは通常、試験官長が作成し、評点が特に低かった、または高かった設問やトピックを含む各評価要素についての情報、および試験での生徒のパフォーマンスについての情報が提供されます。また、今後のセッションで類似の設問やトピックに効果的に取り組むために生徒の準

備を整える方法について、試験官チームから出された一般的な推奨事項も含まれます。科目レポートはまた、その科目の成績区分、および個別の要素の評点区分も示します。

各教師は、科目レポートの情報を自分の教室の文脈にあてはめて考える必要があります。例えば、特定の設問について、全体的には評点が低かったものの、自分が担当している生徒は全員高い評点を獲得したという状況も十分に考えられます。

内部評価のフィードバック

内部評価で教師が付与した評点は、モデレーションの対象となります。内部評価の評価規準について、教師の解釈が IB の試験官の解釈と異なる可能性があります。例えば、教師の採点の方が厳しく、低い評点が付与されているかもしれません。内部評価のモデレーションの目的は、教師の採点基準を調整し、主任試験官またはグローバル基準と一致させることです。教師の採点がグローバル基準とどのように異なっているかを理解できるよう、IB は、教師の評点が許容差を超えていた場合、内部評価の採点についてフィードバックを提供します。これにより、モデレーション係数が適用された理由を教師が理解できるようになります。

学校に提供されるフィードバックは、どのようにしていれば生徒のサンプルの評点がもっと高くなっていたかを説明するものではありません。このフィードバックでは、教師がどの程度適切に各評価規準を採点したか、および評価規準の全体的な適用と、提出された成果物の適性が述べられています。

教科書、ワークショップ、試験に関する注意事項

各 IB コースのカリキュラムは科目の『指導の手引き』に記載され、これを基に評価が設計されます。IB が承認したものを含むあらゆる教科書は、生徒と教師が科目の『指導の手引き』で定められたとおりにコースを完了するための参考資料として使われるものであり、カリキュラムの範囲を定義するものではありません。

したがって、トピック全体またはトピックの一部が、教科書には載っていないものの『指導の手引き』に記載されている場合には、そのトピックに関連する設問が試験で出題される可能性があります。したがって、カリキュラムの範囲を検討する際は、『指導の手引き』のみを参照することがすべての教師に推奨されます。IB は、試験作成プロセスの一環として、一般的に使用されている教科書を確認し、その中に記載されている問題が試験で出題されないようにしています。

同様に、IB が開催するワークショップでのコメントや配布物は、『指導の手引き』の文言をどのように解釈すべきかを理解するヒントとなりますが、『指導の手引き』に定められた内容自体を置き換えるものではありません。『指導の手引き』の改訂は IB によって正式に公開され、その旨が明確に説明されます。

プログラム固有のプロセスの定義

- IB のすべてのプログラムは同じ教育理念に支えられています。そのため IB のすべての評価は、この理念を満たすことができるよう、同じ原則および幅広い実践要綱に従う必要があります。教育的成熟度の段階ごとに児童生徒が次のプログラムへと学習の歩みを進めていくなかで、評価の目的も自然と変わっていきます。
- IB プログラムにおける総括的な外部評価の大まかな目的は、生徒の学力達成度に関する妥当な要約を提供し、進学や就職に向けて進んでいく生徒をサポートすることにあります。
- PYP においても評価は重要な要素の 1 つですが、PYP で総括的な外部評価を行うことは適切ではありません。
- 各プログラムには独自の特徴があり、評価の実践という枠組みにおいて固有のプロセスが求められます。

全プログラムに共通の要素

- IB プログラムでは、幅広く、バランスのとれた、概念的で、関連性の高いカリキュラム、またはカリキュラムの枠組みを提供しています。
- IB の評価はすべて、設計段階においてこの IB 教育の根本的な要素を考慮に入れる必要があります。これらの要素は明示的に評価されないこともありますが、学習と指導に望ましい逆流効果をもたらすことができます。

図 39
各 IB プログラムの連続性

IB 資料『国際バカロレア（IB）の教育とは』（2019 年発行）によれば、各プログラムは概念型の学習を促し、複数の教科において関連性を持ち、学習を統合しカリキュラムに一貫性を与える有力な考えを体系化することを重視します。IB のプログラムは、学習したことの間につながりを見出し、学問領域の間の関係を探究し、個別の教科や科目の枠にとらわれずに世界について学ぶことの重要性を強調しています。また、幅広く、バランスのとれた、概念的で、相互につながりのあるアカデミックな学習活動に触れる場を、さまざまな形で児童生徒に提供します。

- **幅広くバランスのとれた学習内容** — IB の教育では、児童生徒がさまざまな教科にまたがる幅広い内容にアクセスできるようにバランスのとれた教育方法を採用しています。
- **概念型の学習** — 概念型の学習では、各教科や教科横断的な領域において関連性をもつ、幅広く、有力な考えを体系化することを重視します。概念は国や文化の境界にとらわれるものではありません。概念は、学習内容を統合し、カリキュラムに一貫性をもたせます。また、教科学習の理解を深め、複雑な考えに取り組む力を築き、学習内容を新たな文脈に適用するのに役立ちます。

- **相互に関連する学習内容** — IB のカリキュラムの枠組みは、学習の同時並行性（concurrency of learning）を重んじます。プログラムの中で、児童生徒は同時並行的に多くの科目に取り組みます。関連性を見出すことを学び、多岐にわたる分野における知識や経験が相互に関連し合うことについての理解を深めます。各コースのねらいとプログラムの要件は、教科の枠にとらわれず世界について学べる真の機会を提供するよう設定されています。

それぞれの評価を設計するにあたり、教師、学校、IB の評価作成者は、質の良い学習と指導を損なうような課題を作成しないよう、この基本的な目標を振り返る必要があります。評価の狙いは常に、望ましい逆流効果を生み出すことです。

国際的側面

「IB の使命」は、児童生徒が「多様な文化への理解と尊重の精神を通じて、より良い、より平和な世界を築くことに貢献する思いやりに富んだ若者」「人がもつ違いを違いとして理解し、自分と異なる考えの人々にもそれぞれの正しさがあり得ると認めることのできる人」として成長する手助けをすることです（「IB の使命」、2024）。IB プログラムは、多数の国や地域（統治領）で、多数の国籍をもつ児童生徒に提供されています。したがって、IB の学習と指導には国際的な文脈と国際的な視野への焦点の両方が存在し、評価にはその両方が反映されていなければなりません。国際的な視野についての IB のアプローチについて詳しくは、本資料の「国際的な視野と多様な文化への理解」のセクションを参照してください。

「IB の学習者像」

> 今日の教育は、問題解決と意思決定に対する創造的で批判的なアプローチを含んだ考え方を非常に強く重視しています。さらに、コミュニケーションや協働などの取り組み方に加えて、新しいテクノロジーの可能性を認識して利用する能力や、新しいテクノロジーのリスクを実際に回避する能力など、そのような取り組み方に必要なツールも重要視されています。最後に、教育は、能動的で積極的に関与する市民として多面的な世界を生きるための能力にも焦点を合わせています。このような市民は自分が学習したいことや求める学習方法に影響を及ぼし、それが教育者の役割を形づくることになります。
>
> (Schleicher, 2016)

21 世紀型のスキルや能力を分類する方法には、経済協力開発機構（OECD）の 21 世紀型能力、RAND Education、全米研究評議会（NRC）の枠組みなどさまざまなものがありますが、IB では、「IB の学習者像」においてこれらの能力を説明しています。

図 40
「IB の学習者像」

「IB の学習者像」のすべての側面が総括的評価を通した測定に適しているわけではありません。優れた学習、指導、評価はそのような特質の重要性を認識し、たとえそのような特質を測定するように設計されていない場合でも、上記の能力を育成する機会を児童生徒に与えることができます。その例として挙げられるのが、倫理的な（信念に基づいた）調査と実験のアプローチの促進を通して、適切な相互評価をサポートし、予想外の文脈を児童生徒に紹介することです。

関連文献

IB の教育プログラム全体を支える価値観について詳しくは、以下の資料を参照してください。

- IB 資料『国際バカロレア（IB）の教育とは』

- 「IB の学習者像」
- IB 資料『原則から実践へ』（MYP、DP、CP（英語版））

プログラム固有のニーズと解決策

本資料の既出セクションで、評価サイクルを形成するプロセスについて説明しました。この全般的なプロセスは IB のすべてのプログラムにあてはまります。ただし、各プログラムにはそれぞれ個別の目的があり、各プログラムの評価、およびプログラムの証明書が付与される条件については、互いに異なる部分があります。

プログラムごとに異なる評価へのアプローチは、それぞれ大きな重要性をもちます。本セクションでは、各プログラムの評価へのアプローチと、その管理方法について説明します。

ディプロマプログラム（DP）

DP で実施される総括的な外部評価には、以下のような特徴があります。

- 生徒がディプロマを授与されるには、規定された複数の科目を受験しなければならない。
- ディプロマにおける全体の到達度はスコアで表される（最高 45 点）。
- 「コア」要素はスコア換算表に基づいてディプロマの全体スコアに加点される（最大 3 点）。
- ほとんどすべての科目が複数の評価要素をもち、外部評価と内部評価の両方が用いられる。
- ほとんどすべての科目が SL と上級レベル HL で提供される（ディプロマの全体成績への貢献度は同じ）。

DP のねらい

評価の結果は、コースおよびプログラムのねらいと明確に結びついて初めて妥当性をもちます。

DP のねらいは、16 歳～19 歳の生徒を対象に、チャレンジに満ち、国際的な視野に立ち、幅広くバランスのとれた教育体験を提供することです。このねらいを支えるために、2 年間の課程で 6 科目とカリキュラムの中核を成す「コア」（必修要件）を履修することが求められます。DP はまた、目的のある充実した人生を送るために必要な価値観や生活力を育みながら、生徒が大学とその後の高等教育、そして将来選択する職業で必要とされる基本的な学問的能力を身につけることができるよう設計されています。つまり、DP およびその各コースと「コア」要素の評価結果の妥当性は、各生徒の全体的な教育体験に支えられています。

DP の評価結果の適正な用途

IB は、DP コースの成績と全体的なディプロマの点数が以下の用途に使用されるという前提で、評価モデルとカリキュラムを開発しています。

- 大学の入学者選考または就職時の採用選考。
- 大学課程の要件の一部を生徒がすでに満たしているかどうかを決定する。これが認められた場合、授与された資格に対して単位を与える、または特定の科目の履修を免除するなどの措置がとられる可能性がある。
- 特定の言語において、さらなる学習に取り組むための能力と資格があることを示すエビデンスを提供する。

これらの用途を検討することが重要な理由については、本資料の「妥当性」のセクションを参照してください。

DP および各コースの構成

生徒は IB ディプロマを取得するために、6 つの科目、および CAS、「課題論文」（EE：extended essay）、「知の理論」（TOK：theory of knowledge）から成る DP の「コア」を履修しなければなりません。

生徒は、いくつかの科目を SL で、いくつかを HL で履修します。各生徒は、少なくとも 3 科目（最大 4 科目）を HL で履修し、残りの科目を SL で履修します。SL のコースと HL のコースは学習範囲が異なりますが、教科共通の成績評価の説明と評価目標を通して、両コースのパフォーマンスの整合性が確保されます。HL では、生徒はより高度な知識、理解、スキルの発揮が求められます。

少数ながら提供されている学際的なコースは、複数の教科として数えられます。例えば、「環境システムと社会」（ESS：environmental systems and societies）を履修すれば、「個人と社会」および「理科」から 1 科目を履修するという要件が同時に満たされます。

図 41
DP のプログラムモデル

DP の成果の計算

DP の妥当性は、生徒のパフォーマンスの測定方法に反映されています。全体的なディプロマのスコアは、6 つの科目で獲得した成績（1～7 点）を合計し、さらにそれに「コア」

要素（TOK、EE、CAS）のパフォーマンスを表す点数（0〜3 点）を加えることで計算されます。SL と HL の科目は、生徒の最終的な点数を決めるうえで、同じように扱われます。生徒が獲得できる最大スコアは 45 点です。

「コア」のスコア換算表

科目の場合と異なり、TOK と EE では A〜E の 5 段階で成績がつけられます。3 つ目の「コア」要素である CAS には、その性質上、成績は付与されません。

「コア」での成績は 0〜3 点に換算され、ディプロマの全体スコアに加算されます。TOK または EE で E の成績を付与された場合、または CAS の要件を完了しなかった場合は、IB ディプロマを取得することはできません。「コア」の点数を計算する方法を図 42 に示します。

図 42
TOK と EE の点数の付与

		「知の理論」（TOK）				
	付与される成績	A	B	C	D	EまたはN
「課題論文」（EE）	A	3	3	2	2	ディプロマ取得不可
	B	3	2	2	1	ディプロマ取得不可
	C	2	2	1	0	
	D	2	1	0	0	
	EまたはN	ディプロマ取得不可				

IB ディプロマの取得要件

生徒が IB ディプロマを取得するには、以下の条件を満たす必要があります。

- CAS の要件を満たしている。
- 生徒の総合点が 24 点以上である。
- TOK、EE、または要件となる科目に対して「N」（成績なし）が付与されていない。
- TOK と EE の両方で少なくとも成績 D が付与されている。
- どの科目／レベルでも成績 1 が付与されていない。
- 成績 2 が 3 回以上付与されていない（SL または HL）。
- 成績 3 以下が 4 回以上付与されていない（SL または HL）。
- HL 科目の獲得点数が 12 点以上（4 つの HL 科目に登録している生徒の場合、上位 3 科目の成績を合計する）。
- SL 科目の獲得点数が 9 点以上（2 つの SL 科目に登録している生徒の場合、SL で少なくとも 5 点を獲得しなければならない）。

バイリンガルディプロマ

以下の規準のうち**1 つ以上**を満たす生徒は、標準的な IB ディプロマの代わりにバイリンガルディプロマを取得できます。

- 「言語と文学」から選択した 2 つの言語を修了し、両方の言語で 3 以上の成績を獲得している。

- 「言語と文学」で生徒が指定した学習言語とは異なる言語で、「個人と社会」または「理科」のうち 1 科目を修了している。この場合、「言語と文学」と「個人と社会」または「理科」から選択した科目の両方で、3 以上の成績を獲得しなければなりません。

上記の条件が満たされていれば、パイロット科目や学際的な科目もバイリンガルディプロマの授与に必要な科目とみなすことができます。

以下はバイリンガルディプロマの取得条件を満たすための科目とはみなされません。

- EE
- 学校独自シラバス（SBS：school-based syllabus）科目
- IB ディプロマに必要な 6 科目以外に生徒が受講した科目（「追加の科目」と呼ばれる）

すべての受験者群で同じ基準を維持する

DP の科目の受験者群は、その規模が科目によって大きく異なり、受験者が 1 人のこともあれば、40000 人以上になることもあります。すべての科目は等しく重要ですが、実践的な観点から、受験者数が多い科目と少ない科目ではプロセスが若干異なります。

試験官の質を維持する

試験官の数が非常に少ない科目では、本資料の「標準化」のセクションで説明した品質モデル（練習用スクリプト、認定用スクリプト、シードスクリプト）の採用が適切ではない場合があります。

すべての試験官が標準化の議論に参加できる場合、各試験官がすでに基準の設定に寄与しているため、練習用スクリプトや認定用スクリプトを提供する必要はなくなります。この場合でも、試験官が期待された基準に沿って採点していることを確認するためにシードスクリプトは使用されます。ただし通常は、要否にかかわらず完全な品質モデルが作成されます。

重要なことは、科目の受験者が少ない場合でも採点の質を維持することです。

成績：完全な成績付与（バーチャルまたは対面）

試験官長または主任試験官の下に数名の試験官が配置される科目では、本資料の「成績付与と集約」のセクションに記載された実践に従って、正式な会議が開かれます。試験官を補佐し、プロセスの品質確認チェックを実施するため、IB のサブジェクトマネージャーが関与します。この会議は、対面でもバーチャルでもかまいません。

成績：受験者が少ない科目のガイド付き成績付与（バーチャル）

各要素を担当する試験官の数が非常に少ない場合は、上級試験官がバーチャルで話し合いを行います。IB のサブジェクトマネージャーが、成績の付与を支える幅広いエビデンスを提供するとともに、品質確認のためにプロセスのモニタリングを行います。

成績：受験者が少ない科目の標準的な成績付与

受験者の数が少なく、年度ごとの統計的なエビデンスがあまり意味をなさない場合、試験官は判断に基づく成績区分において、サンプル成果物ではなく、評価のために提出されたすべての成果物を確認するよう求められます。

IB のサブジェクトマネージャーは、要請に応じてこの会議をサポートし、定期的に会議を監督するのではなく、無作為の品質チェックを実施します。

受験者が少ない科目すべて（ガイド付きおよび標準的な成績付与）で試験官をサポートするために、同一教科内の異なる科目の試験官同士が話し合いをもつことが推奨されます。これにより、それぞれの成績の意味について、教科内で共通の理解を確立することができます。

最終承認

試験を受けた生徒の数にかかわらず、成績付与プロセスの最終段階は、試験官長または主任試験官が IB の上級職員に対して提言を行うことであり、IB の上級職員は、この提言を裏づけるエビデンスが有効であることを確認します。

すべての科目が、その受験者数を問わず同じ方法で精査されます。

関連文献

DP についての詳しい情報は、以下のリソースをご覧ください。

- IB 資料『DP：原則から実践へ』
- IB 資料『ディプロマプログラム（DP）における評価の手順』
- IB 資料『IB ワールドスクールのための規則』
- 各科目の『指導の手引き』（プログラム・リソース・センターの科目ページに掲載）
- IB 資料『教師用参考資料』（プログラム・リソース・センターの科目ページに掲載）
- IB 資料『ディプロマプログラム（DP）における評価に基づく指導と学習』

キャリア関連プログラム（CP）

CP の評価には、以下のような特徴があります。

- 生徒が IB の CP 修了証を取得するには、所定の要件を満たす必要がある。
- IB の CP 修了証に関連する全体スコアは存在しない。
- DP と共通のコースを履修する CP の生徒は、DP の生徒と一緒に評価される。
- CP の枠組みでは、IB によって提供されず、成績も付与されない「キャリア関連学習」（CRS：career-related studies）が求められる。

CP のねらい

評価の結果は、コースおよびプログラムのねらいと明確に結びついて初めて妥当性をもちます。

CP ならでは特徴は、キャリアに特化した学習者としての生徒の育成を支えるということです。進学であれ就職であれ、自分が希望する進路において役立つ、生涯にわたって活用できる転移可能なスキルの発展を目指します。

CP は、以下の点において生徒をサポートします。

- DP コースと「キャリア関連学習」を通して、幅広い能力を発展させ、特定の「知識の領域」における理解を深める。
- さまざまな文脈において知識とスキルを獲得する、または知識とスキルを向上するための柔軟なストラテジーを策定する。
- さまざまなものの見方、機会、課題に向き合える生涯学習者となるために必要な姿勢や習慣を身につける。
- 変化を続ける世界に効果的に参画するための力を身につける。
- 建設的な貢献を行う能力と意志を育む学習に積極的に参加する。

CP の評価結果の適正な用途

IB は、CP コースの成績と CP 修了証が以下の用途に使用されるという前提で、評価モデルとカリキュラムを開発しています。

- 就職時の採用選考、およびインターン等の雇用プログラムの選考プロセス。
- 適切な専門分野で研究を続けるための選考プロセス。
- 大学の入学者選考。
- 大学課程の要件を生徒がすでに満たしているかどうかを決定する（追加単位の付与、または特定の科目の履修免除）。

生徒が特定の（使用）言語で評価を受けた場合、その言語で当該科目や専門分野をさらに学習できることを示すエビデンスとなります。

これらの用途を検討することが重要な理由について詳しくは、本資料の「妥当性」のセクションを参照してください。

CP の構成

CP は 3 つの部分からなる教育的な枠組みです。以下の要素で構成されています。

- 2 つ以上の DP 科目（SL または HL）
- CP の「コア」要素
- 「キャリア関連学習」

図 43
CP プログラムのモデル

CP の「コア」要素

CP の「コア」の 4 つの要素は、その性質、学習成果、意図的なつながりという意味で互いに関連しています。「コア」の全体的なねらいは以下のとおりです。

- 「IB の学習者像」および国際的な視野を育てることを通して、「IB の使命」に根ざしたプログラムを提供する。
- 枠組みのすべての要素をつなげることで、DP コースと「キャリア関連学習」を文脈化し、その効果を高める。
- 長期にわたって個人的、学問的、専門的な知識、スキル、姿勢の育成を促す。
- 学習の反復性と相互関連性、および個人とコミュニティーの健やかさを維持するうえで学習が果たす重要な役割への理解を反映する。

CP では以下の 4 つの「コア」要素をすべて完了しなければなりません。

- 「コミュニティー活動」
- 「言語と文化の学習」
- 「パーソナルスキルと職業的スキル」
- 「振り返りプロジェクト」

「振り返りプロジェクト」は評価の対象です。つまり、完成したプロジェクトと振り返りが採点され、生徒はその評価要素に対して IB から最終成績を授与されます。その他の「コア」要素については、各要素が問題なく完了したことを学校が IB に報告する必要がありますが、IB が学校の評価を確認することはありません。CP を実施する学校におけるプロセスと実践は、学校への評価訪問の中で確認されます。

CP の一部を構成する DP コース

各生徒が、DP との共通科目を少なくとも 2 コース履修します。CP の生徒は、DP の生徒と同じ評価プロセスに組み込まれます。CP の生徒に特化した個別の試験や成績付与のプロセスはありません。

DP と CP では一部のコースが共通しているものの、同時に両方のプログラムに登録することはできません。プログラムごとに幅広い要件が設定されていることを考えると、2 つのプログラムを並行して履修することは不可能です。

「キャリア関連学習」

CP の「キャリア関連学習」の結果を IB が評価したり、品質確認を行ったりすることはありません。唯一の要件は、生徒が「キャリア関連学習」を完了したことを IB に報告するということのみです。DP コースを CP の「キャリア関連学習」の一部に含めることは適切ではありません。

CP の成果の計算

IB の CP 修了証に関連する全体スコアは存在しません。

IB の CP 修了証は、次の要件をすべて満たした生徒に授与されます。

- 指定された「キャリア関連学習」を修了したことを学校が確認している。
- 少なくとも 2 つの DP コースで 3 以上の成績を取得している。
- 「振り返りプロジェクト」で D 以上の評価を取得している。
- 「コミュニティー活動」「言語と文化の学習」「振り返りプロジェクト」「パーソナルスキルと職業的スキル」の要件を満たしていることを学校が確認している。
- 資格授与委員会から学問的不正行為に対する罰則を受けていない。

CP の成績と結果は、DP と同じ資格授与委員会によって確認されます。

バイリンガル CP 修了証

以下の規準のうち **1 つ以上**を満たす生徒は、CP 修了証の代わりにバイリンガル修了証を取得できます。

- 「言語と文学」から選択した 2 つの DP の言語コースを修了し、両方で 3 以上の成績を獲得している。
- 「言語と文学」の DP 言語コースを 1 つ修了し、「言語と文学」を履修した言語とは異なる使用言語で「個人と社会」または「理科」の DP コースを修了している。両方のコースで 3 以上の成績を獲得していなければならない。

基準の維持および成果物が確認できない場合の手順

評点がつけられない場合の手順および受験者数にかかわらず同じ基準を維持することについては、「振り返りプロジェクト」および DP と共通のすべてのコースで DP と同じアプローチがとられます。

成果物が確認できない場合、および評点がつけられない場合について詳しくは、本資料の「評点がつけられない場合」のセクションを参照してください。

関連文献

CP についての詳しい情報は、以下のリソースをご覧ください。

- IB 資料『Career-related Programme: From principles into practice』
- IB 資料『Career-related Programme Assessment procedures』
- IB 資料（英語版）『Overview of the Career-related Programme（キャリア関連プログラム（CP）の概要）』
- IB 資料（英語版）『Reflective project guide（「振り返りプロジェクト」指導の手引き）』
- IB 資料（英語版）『Language and cultural studies guide（「言語と文化の学習」指導の手引き）』
- IB 資料（英語版）『Personal and professional skills guide（「パーソナルスキルと職業的スキル」指導の手引き）』
- IB 資料（英語版）『Community engagement guide（「コミュニティー活動」指導の手引き）』

中等教育プログラム（MYP）

MYP 修了証は、生徒が MYP のすべての側面を修了したことを示す証明書です。MYP 修了証を取得するには、以下の要件を満たす必要があります。

- MYP e アセスメントのいずれにおいても成績 1 か成績 2 がなく、合計が 28 点以である。
- 推奨される 2 年間（少なくとも 1 年間）にわたってプログラムを履修し、第 5 年次の要件を完了している。
- 最低 5 科目（異なる教科から選択）、最高 8 科目（8 つの教科から選択）で内部評価と試験を受けている（必須科目が含まれていなければならない）。
- 「芸術」「保健体育」「デザイン」の少なくとも 1 科目で e ポートフォリオを完成させている。
- 学際的な学習でコンピューターを用いた試験を受けている。
- 「パーソナルプロジェクト」を完成させ提出している。
- 「コミュニティー活動」に関する IB の最低要件を完了している。

MYP のねらい

評価の結果は、コースおよびプログラムのねらいと明確に結びついて初めて妥当性をもちます。

> IB の MYP は、11 歳から 16 歳の生徒を対象に、やりがいのある学習活動を提供し、人生に必要なスキルを育成する、一貫的かつ包括的カリキュラムの枠組みとして構築されてきました。 この時期は、若者の成長発達における臨界期であり、学校での成功は、個人的、社会的、情緒面での健やかさに深く関連しています。アイデンティティーが確立し自尊心が育つこの重要な時期において、MYP は生徒に意欲を与えるとともに、彼らが教室や学校の枠をこえた人生において成功するうえでの助けとなります。MYP は、生徒が個々の強みを足がかりに、あまり得意としない科目での挑戦をも受け入れることができるようなプログラムになっています。また、生徒が自分の可能性を広げ、学習における自分自身の好みや傾向を探り、身の丈にあったリスクを負うことにも挑戦し、確固とした自分だけのアイデンティティー意識と向き合い、それを育む機会を提供します。
>
> （IB 資料『MYP：原則から実践へ』（2014 年版）、p. 3）。

MYP では、MYP の教科目標と採点規準の間に明確な整合性があります。すべての MYP 教科で、4 つの目標に対応する 4 つの評価規準が定められています。各規準が、最終的な成績に同じ割合で寄与します。

e アセスメントにグローバルな文脈を組み込む

> MYP において、学習の文脈は、実世界の背景や出来事、状況などであるか、それらをモデルとするべきです。MYP における学習の文脈はグローバルな文脈から選択され、国際的な視野の育成とプログラムの中でのグローバルな取り組みを促進します。（中略）MYP は、指導と学習に対して 6 つのグローバルな文脈を特定しています。それらの文脈は PYP の教科の枠をこえた学習テーマに基づいて展開します。
>
> （『MYP：原則から実践へ』（2014 年版）、p. 23）。

各試験セッションは、『MYP：原則から実践へ』のリストから選択された特定のグローバルな文脈と探究の中に位置づけられ、それを土台とします。

各科目のコンピューターを用いた試験では、約 3 分の 1 の課題が所定のグローバルな文脈に関連するもの、所定のグローバルな文脈から着想を得ているもの、または所定のグローバルな文脈に由来するものとなります。学際的なコンピューターを用いた試験は、そのすべてが所定のグローバルな文脈に着想を得たものとなります。

「言語の習得」「芸術」「デザイン」「保健体育」の一部記入済みの単元プランナーは、所定のグローバルな文脈を参照して作成されます。

MYP の評価結果の適正な用途

IB は、MYP コースの成績と修了証のスコアが以下の用途に使用されるという前提で、評価モデルとカリキュラムを開発しています。

- 進学または就職のための選考プロセス。
- 教育を続ける生徒に前向きなフィードバックを与え、個人的な強みを示す。

生徒が特定の（使用）言語で評価を受けた場合、その言語で当該科目をさらに学習できることを示すエビデンスとなります（例：フランス語、スペイン語、または英語で教育を行う学校への進学）。

これらを検討することが重要な理由について詳しくは、本資料の「妥当性」のセクションを参照してください。

MYP の構造

図 44
MYP のモデル

MYP の成果の計算

MYP の第 5 年次の終わりに、生徒は IB の外部評価を受けることができます。この評価の結果は、MYP 科目別履修証という資料に記録されます。さらに、生徒が希望する場合は、MYP 修了証の授与につながる評価を受けることもできます。

学校が MYP 履修記録を発行することもできます。これは、少なくとも 2 年間はプログラムの科目を履修し、第 3 年次または第 4 年次における要件を満たした生徒が対象です。また、IB の評価に一切登録していないものとします。MYP 履修記録は学校が発行する文書であり、IB が認証したものではありません。

MYP 修了証を取得するには、プログラムの最終学年を履修し（2 年間履修することが望ましい）、以下の要件を満たしている必要があります。

- 「言語と文学」「言語の習得」（または「言語と文学」から追加でもう 1 科目）、「個人と社会」「数学」「理科」、および「芸術」「保健体育」「デザイン」から選択した 1 科目からなる合計 6 つの科目で、コンピューターを用いた試験または e ポートフォリオを完了している。
- 上記に列挙した 6 つの科目それぞれで、少なくとも成績 3 を獲得している。
- 学際的なコンピューターを用いた試験を完了し、少なくとも成績 3 を獲得している。

- 「パーソナルプロジェクト」を完了し、少なくとも成績 3 を獲得している。
- 合計 28 点を獲得している。
- 「コミュニティー活動」に対する学校の期待事項を満たしている。

MYP バイリンガル修了証を取得するには、さらに、次のいずれかのコンピューターを用いた試験で規定以上の成績を修める必要があります。

- 「言語と文学」の 2 つ目の科目（「言語の習得」の 1 科目の代わりに履修）
- 「理科」「個人と社会」または学際的な科目のうち少なくとも 1 科目（「言語と文学」コースにおいて選択した言語以外の言語で受験）。

総括的外部評価の実施（MYP e アセスメント）

任意の e アセスメントでは、以下の 2 つの方法で生徒の知識と能力を評価します。

- e ポートフォリオ：「言語の習得」「芸術」「デザイン」「保健体育」において評価のために提出される成果物。モデレーションを通して、グローバル基準が一貫して適用されていることを確認。
- コンピューターを用いた試験（2 時間）：「言語と文学」「言語の習得」「個人と社会」「理科」「数学」「学際的な学習」のコースで実施。

これに加えて、「パーソナルプロジェクト」を電子形式で IB に提出し、モデレーションを受けます。MYP 第 5 年次の生徒は必ず「パーソナルプロジェクト」の e アセスメントを受けなければなりません。他の e アセスメントは任意です。

試験の基本計画

IB は、e アセスメントについて学校が明確に理解できるよう、試験の基本計画を発行します。この基本計画を精査することで、教師と生徒は、MYP e アセスメントの特徴と目的を理解できるようになります。コンピューターを用いた試験の準備を進めるうえで役立つ資料であり、生徒が教科の規準と各教科の評価方法に集中して対応できるようにします。基本計画には常に 4 つの規準が含まれ、各規準が同じ比重を占めます。

IB は、どのセッションにおいても、試験が基本計画から 3 点以上逸脱することがないようにします。

e ポートフォリオと一部記入済みの単元プランナー

e ポートフォリオは、時間をかけて取り組むコースワーク課題（成果物）やパフォーマンスを評価するための手段です。このような課題はその性質上、試験で評価することが容易ではありません。e ポートフォリオの基盤となるのが一部記入済みの単元プランナーです。教師はこれをガイドとして使用することで、地域の文脈に柔軟に対応しながら、公正で有意義な判断をするための適切な生徒のエビデンスを生み出せるようになります。一部記入済みの単元プランナーは、セッションごとに新しく提供されます。

単元プランナーは、教師が設定する課題が、MYP の成績全体にわたって生徒のエビデンスを示すものになっていることを確認する役割を果たします。教師が設定する課題には、難易度が低すぎるまたは高すぎることによって生徒が不利益を被るリスクが伴います。モデ

レーションにおいて IB は、評価のために提出された成果物のみに基づいて成績を付与します。教師が設定した課題が成績評価の説明の一部にしか対応していない場合、範囲外の成績を付与することはできません。

評価の一元化 ― 生徒の負担を管理する

試験を受け、コースワークに取り組むことは、生徒に多くを求めるストレスの多い作業です。また、指導時間が削られてしまうこともあります。MYP の第 5 年次の生徒については、あらゆる点からみて、総括的評価の量をできる限り減らすことが適切だと考えます。これによって、生徒が自分の到達度を実証する機会が 1 回しか与えられないという問題が生まれますが、IB では、生徒の心身の健やかさを確保するために、この問題を甘んじて受け入れています。

成績がつけられない場合の対応

評価要素を 1 つに絞ることで生徒の作業負担を管理することができますが、MYP では、利用できるエビデンスが制限されてしまうというデメリットが生まれます。これは、成績がつけられない場合の手順に対する IB の信頼度が、他のプログラムにおいて評点がつけられない場合の手順に対する信頼度に比べて低くなることを意味します。したがって、本手順は例外的な場合にのみ使用すべきです。

成績がつけられない場合の手順について詳しくは、本資料の「成績がつけられない場合の手順」を参照してください。

関連文献

MYP についての詳しい情報は、以下のリソースをご覧ください。

- IB 資料『MYP：原則から実践へ』
- IB 資料（英語版）『Guide to the MYP exam session（MYP 試験セッションに関する手引き）』
- MYP の科目の『指導の手引き』
- MYP「プロジェクト」の指導の手引き
- IB 資料『Middle Years Programme assessment procedures』
- IB 資料（英語版）『MYP on-screen examinations: IT requirements and school responsibilities（MYP のコンピューターを用いた試験：IT 要件と学校の責任）』
- IB 資料（英語版）『MYP on-screen familiarization（MYP のコンピューターを用いた試験の参考資料）』

初等教育プログラム（PYP）

PYP の評価には、以下のような特徴があります。

- 評価に際しては、教師と児童が協働して学習についてのモニタリング、記録、測定、報告、および調整を行う。
- 評価の文化を育むためには、学習コミュニティーに属するすべての人が評価する能力をつけなければならない。
- IB の総括的外部評価の要件はない。

PYP のねらい

評価の結果は、コースおよびプログラムのねらいと明確に結びついて初めて妥当性をもちます。

PYP カリキュラムの枠組みは、従来の教科をつなぎ、すべての教科にわたり、かつ教科を超越するような学習を児童が経験できるよう、「教科の枠をこえた学習」を中心に置きます。

構成主義および社会構成主義的な学習理論に基づく PYP カリキュラムの枠組みは、PYP の児童は自らの学習におけるエージェント（エージェンシーをもつ人）であり、学習プロセスのパートナーであるという前提から始まります。児童は、自分自身、他者、周囲の世界について探究し、質問し、疑問をもち、理論化する潜在能力を生まれながらにしてもっています。

児童がどのように学ぶかということに関するこの理解は、探究型で概念に基づく学習と指導の教科の枠をこえたモデルの基礎となります。探究プログラムに取り組み、自らの学習を振り返ることを通して、PYP の学習者は世界と関わり、人々と地球の健やかさと健全性のために行動を起こすための知識、スキル、姿勢を身につけます。

PYP の構造

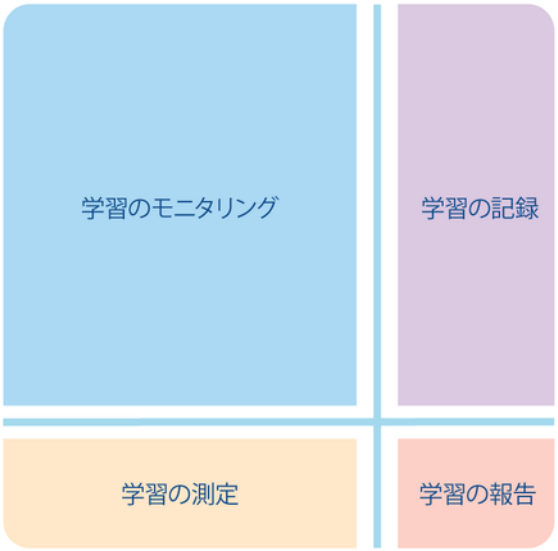
図 45
PYP のモデル

PYP における評価

評価は、科目固有の知識とスキルの獲得、概念的理解、「学習のアプローチ」の発展、および「IB の学習者像」の人物像を通じて、思慮深く効果的に児童をサポートするという PYP の目標において重要な役割を果たします。

PYP の評価には、学習のモニタリング、記録、測定、報告という 4 つの領域があります。各領域には独自の機能がありますが、いずれも学習と指導に情報をもたらすエビデンスを提供することを目的としています。評価の 4 つの領域は、同じ重みで扱われているわけではありませんが、それぞれに重要性と価値があります。PYP では、学習者の行動につながるフィードバックを提供するうえで欠くことのできない学習のモニタリングと記録に重きを置いています。

図 46
PYP における評価の側面



PYP における評価は、以下の前提の上に成り立っています。

- 評価とは、教師と児童の間、および児童同士の間で起こる継続的な協働のプロセスであり、指導を導くための学習のエビデンスを集め、分析し、振り返り、それを行動に移すことを目的としています。
- 児童の学習成果物の評価には、児童が多様かつ複雑で洗練された方法で各自の学習体験を理解することを踏まえ、幅広い方法を活用する必要があります。
- 児童は、自分の学習の評価と振り返りに能動的に関与し、他の児童や教師からのフィードバックに基づいて行動します。これが、学習の次のステップへのフィードフォワードをもたらします。
- 学習目標と成功規準は、協働で構築し、明確に表現して伝えるようにします。
- 学習の結果とプロセスの両方が評価の対象となります。
- 評価の設計は、逆向き設計であると同時に、前向き設計でもあります。
- 評価の文化を育むためには、学習コミュニティーに属するすべての人が評価する能力をつけなければなりません。

関連文献

PYP についての詳細は、プログラム・リソース・センターにある IB 資料『PYP：原則から実践へ』を参照してください。

内部評価のモデレーション：詳細

モデレーションは、すべての学校で一貫した基準が適用されていることを確認するために、MYP、DP、CP の内部評価を対象に実施されます。モデレーションの結果、教師の採点結果が上下に変動することもあれば、そのまま変更されないこともあります。モデレーションの目的は、生徒の成果物を採点する際に、教師が評価規準をどの程度正確かつ一貫して適用しているかを確認することです。

サンプリング

学校から内部評価の評点が提出された後、IB はモデレーションサンプルとなる生徒を選定します。サンプルは、学校の評点範囲全体を適切に表せるよう慎重に選定します。モデレーションのサンプル数は、その科目の評価を受けた生徒数に応じて、10 件、8 件、5 件、4 件以下のいずれかとなります。サンプルの中から、サンプルの評点範囲を代表する 3 人の生徒が最初のサンプルに割り当てられます。これは、教師の採点がグローバル基準の許容差内に収まっているかを決定するために使われます。許容差に収まっていない場合、残りのサンプルも試験官によって採点されます（「拡張サンプル」と呼ばれる）。IB の傾向として、満点を獲得した生徒の成果物をモデレーション用サンプルに選定することはあまりありません。これは、評点範囲の上の方にいる生徒の点数が、モデレーションによって上方修正される可能性を残しておくためです。

コースに登録した生徒の数が多いために履修者が複数のクラスに分けられ、複数の教師が内部評価を行う場合、すべての教師が同じ内部評価を実施し、規準を適用する方法を教師同士で事前に標準化することが求められます。学校が提出するのは 1 つのモデレーションサンプルのみです。これにはおそらく、複数の教師が採点した成果物が含まれることになります。ただし、1 つの学校内で同じ科目を異なる使用言語で指導している場合、言語ごとに別々のモデレーションサンプルを提出する必要があります。

モデレーション係数の定義

すべての内部評価要素は、評価規準またはマークバンドを適用して採点され、ほとんどの場合、評価のために提出された成果物の文脈やプロセスについて、教師は試験官よりも格段に多くの情報を有します。この理由により、内部評価要素のモデレーションを行う試験官は、教師が付与した評点を無視して成果物を再採点するのではなく、教師が付与した採点が適切かどうかを判断するように求められます。教師の評点が変更されるのは、その評点が不適切であり、グローバルな採点基準に沿っていないとモデレーターが確信をもって判断したときのみです。

教師の採点サンプルが試験官によるモデレーションを受けた後、2 つの評点群を統計的に比較し、必要だと判断された場合には、その教師が（その評価要素に対して）学校のすべての生徒に付与した評点に調整が加えられます。

教師の採点が一貫して基準よりも低い、または高い場合は、その教師の評点に一律で同じ調整が加えられます。

表 8 に示す例では、教師が生徒の成果物に対して基準よりも低い評点を付与しています（平均 5 点）。そのため、その教師のすべての採点に対して、5 点が加えられます。

表 8

評点が一貫して基準を下回る場合

教師の評点	試験官の評点	教師の評点と試験官の評点の差	モデレーション後の最終評点
15	20	5	20
11	17	6	16
10	14	4	15
8	12	4	13
4	10	6	9

評点範囲の上方または下方で採点基準が低くなっている、または高くなっている場合は、教師の評点範囲全体で複数の異なる調整が加えられることがあります。

表 9 に、教師がグローバル基準に対して一貫性のない採点を行っている例を示します。評点範囲の上方では採点が甘くなっているのに対し、評点範囲の下方では採点が厳しくなっています。そのため、教師の評点に対して、評点範囲全体で異なる調整が加えられます。

表 9

採点に一貫性がない場合

教師の評点	試験官の評点	教師の評点と試験官の評点の差	モデレーション後の最終評点
37	35	-2	35
25	22	-3	24
17	18	1	17
12	14	2	13
6	8	2	8

教師がすでに正しい基準に沿って採点している場合、調整は加えられません。

表 10 に示す教師の評点はすべてグローバル基準の許容差内に収まっているため、すべての生徒に教師がつけた評点が付与されます。

表 10
正しい基準に沿って採点されている場合

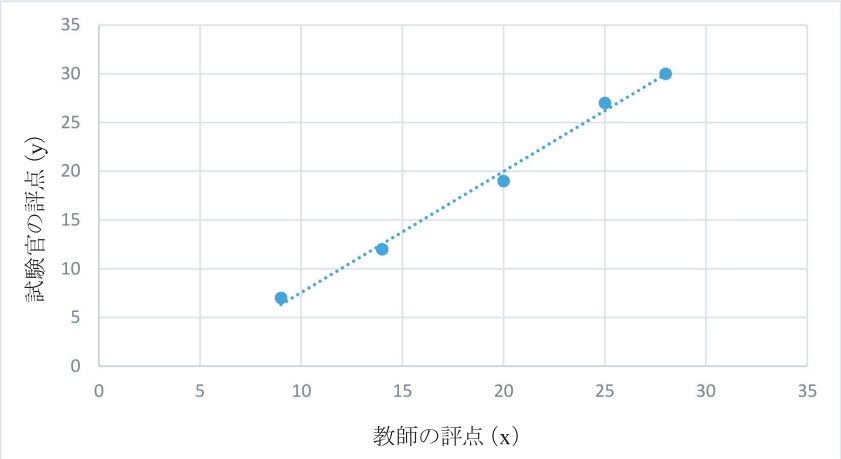
教師の評点	試験官の評点	教師の評点と試験官の評点の差	モデレーション後の最終評点
32	31	-1	32
29	30	1	29
8	8	0	8

線形回帰

はじめに、各モデレーションサンプルのデータを分析し、教師の採点が許容差内にあるかどうかを決定します。許容差内に収まらない場合、サンプルに見られる全体的な傾向に合わせて、その教師の採点すべてに調整が加えられます。この手法は「線形回帰」と呼ばれるもので、教師と試験官が付与したサンプルの評点から導き出されるデータポイント群に最もよく適合する直線を計算で求めます。線形回帰の例と、対応する評点群を図 47 に示します。

図 47 に示すモデレーションの回帰直線は、教師が評点範囲の上方では厳しめに、評点範囲の下方では甘めに採点していることを示しています。各データポイントは、1 つのサンプル成果物に対する教師の評点と試験官の評点のペアを表しています。この継続的な直線を使って、教師の評点がモデレーション済みの評点に換算されます。

図 47
線形回帰の例と対応する評点群



教師の評点	試験官の評点	モデレーション後の最終評点
28	30	30
25	27	26
20	19	20
14	12	12
9	7	6

サンプルデータから計算される回帰直線の式は、教師が付与した各評点（x）を、試験官がその生徒に対して付与すると思われる平均の相当点（y）に換算するために使用できます。1 つのサンプルから大規模な評点群を推定するこのようなモデレーション調整は、採点結果に見られる大まかな傾向を反映することしかできません。モデレーションの目的は、全体的な視点から、生徒の評点をより適切なレベルに調整できるようにすることです。これはベストフィットのアプローチであり、すべての生徒がモデレーション済みの評点として試験官の評点を受け取るわけではありません。

モデレーションの失敗

ベストフィットのアプローチに基づく線形回帰の直線（計算されたモデレーション係数）を、ある教師のすべての採点結果に適用する前に、その直線が一定の条件を満たしていることを確認するための自動チェックが行われます。場合によっては、提出された成果物のサンプルを使ってモデレーション係数を計算できないこともあります。統計的尺度の 1 つに相関係数があります（積率相関係数が使用されます）。これは、教師の採点と試験官の採点の関係の一貫性を測定するものです。

相関係数が 1 の場合、図 48 に示すように、採点結果および生徒を最高から最低まで並べた場合のランクづけ（必ずしも同じ評点である必要はない）の関係が完全に一致していることを示します。

図 48
相関係数 1

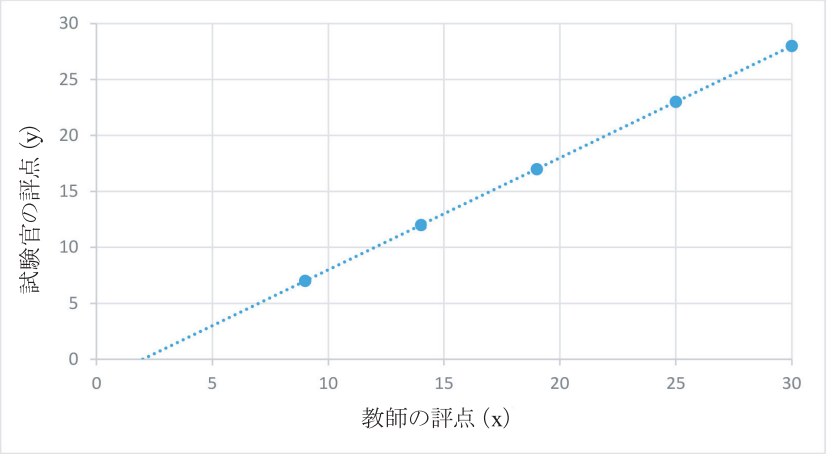
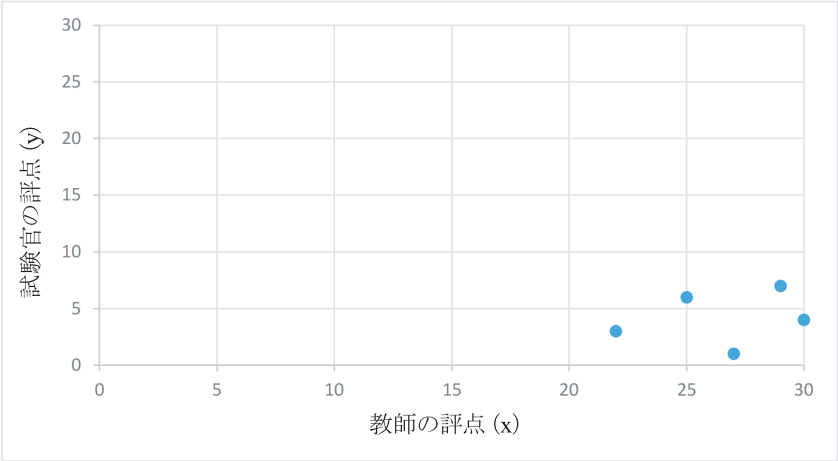


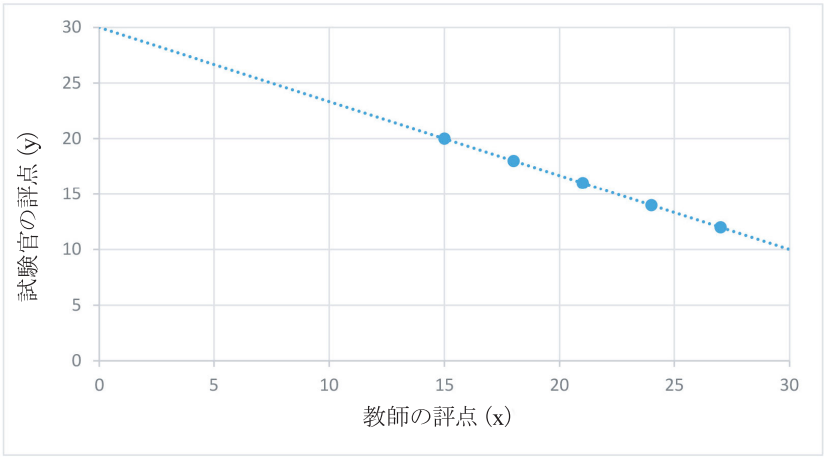
図 49 に示すように、係数 0 はまったく相関がないことを示しています。

図 49
相関係数 0



相関係数が-1 の場合、生徒の成果物の相関的な質について教師と試験官の見解がまったく正反対であり、図 50 に示すように生徒のランクづけが反対になることを示します。

図 50
相関係数 - 1



計算されたモデレーション係数が許容されるためには、相関係数が少なくとも 0.85 となる必要があります。これは、教師と試験官の間の高い一致度を示します。ただし、相関係数が高いからといって、それだけでモデレーション係数が適切であるとはいえません。追加のチェックによって、回帰直線の傾き（勾配）が 0.5～1.5 であることが確認されます。傾きが小さすぎる（浅すぎる）場合、教師の評点が広範囲に広がりすぎていることを示します。つまり、採点結果に一貫性がある場合でも、相対的に質の悪い成果物に評点がほとんど付与されず、質の高い成果物に過度に多くの評点が付与されているということです。このケースでは、試験官は教師の評点範囲を大きく狭める必要がありました。傾きが 1.5 を超える場合、直線の勾配が過度に大きく、先ほどとは正反対の状態となります（評価のために提出された成果物のうち、質の低いものと質の高いものを十分に区別できておらず、付与された評点範囲を試験官が広げる必要があった）。

相関関数が 0.85 未満の場合、または傾きが 0.5～1.5 の範囲に入っていない場合、サンプルは自動モデレーションチェックを通過することができません。

モデレーション失敗となった学校のサンプルはすべて、IB の評価担当職員が個別に確認し、元のデータを慎重に検討したうえで以下のいずれかの判断を下します。

- 計算された回帰直線は、その教師の評点範囲に対しては適切といえる。
- その教師の評点範囲に対して適切とされる他のモデレーション調整を加える。
- 傾向を明確に理解するため、追加のサンプルデータを求める。
- 教師の採点をすべてやり直すため、残りの生徒の成果物の提出を求める。

IB がモデレーションの失敗を解決できるよう、試験セッションが完全に終了するまではすべての生徒の内部評価の成果物を保管しておく必要があります。

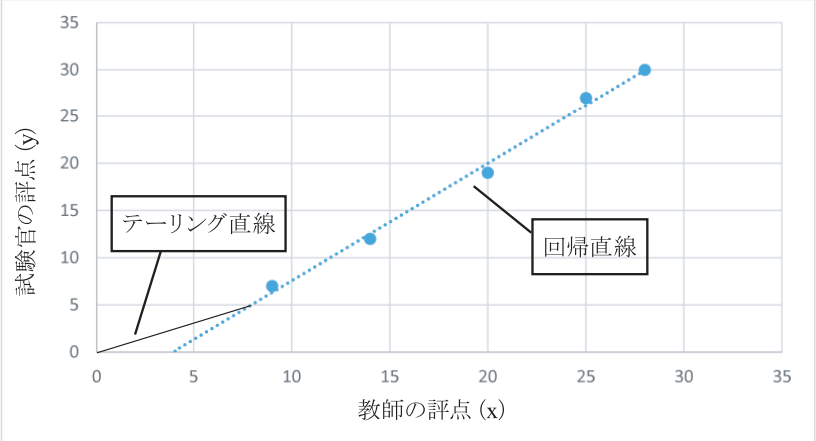
直線モデルの修正

モデレーションに使われる直線モデルは、「テーリング」を使ってある程度修正されます。直線によるモデレーション調整は、どの生徒にも 0 点を付与することができなくなるという意味で、評点範囲の両端部において不適切な効果をもたらすことがあります。成果物の中に評点を獲得するに値する内容が一切書かれていない場合、本来であればその生徒の成果物は 0 点となるべきですが、モデレーションによってこのような成果物に数点が付与されることがあります。また、質は低いものの数点は獲得できるであろう成果物が、モデレーションの結果 0 点になってしまう場合もあります。

この問題を解決するため、評点範囲の下方 20%に位置する評点に「テーリング」という処理を施すことができます。この端部においては、算出された回帰直線は修正され、回帰直線から最小値の座標へとつながる新しい「テーリング」直線に置き換えられます。なお、回帰直線のテーリングは上端部には適用されません。これは、本来その評点に値しない生徒の成果物に対して教師が最高点を付与するケースが、よく見られるためです。

生徒の評点が下限値に向けて下方修正されないようにするための回帰直線のテーリング処理を図 51 のグラフに示します。

図 51
回帰直線のテーリング処理



テーリング処理によって、元々 0 点だった成果物のみが、モデレーション後に 0 点となるようにすることができます。これにより、数点を獲得できるはずの成果物に 0 点が付与されることを防ぐとともに、0 点にすべき成果物に数点が付与されるという事態も回避されます。これは、教師の採点結果が自動モデレーションプロセスを通過したことを前提としています。教師の採点結果がモデレーション失敗となり、自動モデレーションが適用されなかった場合、テーリング処理は実行されません。

参考文献

Ananiadou, K., & Claro, M. (2009). 21st century skills and competences for new millennium learners in OECD countries. OECD Education Working Papers (No. 41). OECD.

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman.

Baird, J. A., Cresswell, M., & Newton, P. (2000). Would the real gold standard please step forward? Research Papers in Education, 15(2), 213-229. https://doi.org/10.1080/026715200402506

Biggs, J. (1996). Enhancing teaching through constructive alignment. Higher Education, 32(3), 347-364. https://www.jstor.org/stable/3448076

Black, P. (1999). Assessment, learning theories and testing systems. In P. Murphy (Ed.), Learners, learning & assessment (pp. 118-134). Paul Chapman Publishing in association with The Open University.

Bloom, B. S. (Ed.), Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: Handbook 1: Cognitive domain. Longman.

Broadfoot, P. M. (1996). Education, assessment and society: A sociological analysis. Open University Press.

Brown, R. (2002). Cultural dimensions of national and international educational assessment. In M. Hayden, J. Thompson, & G. Walker (Eds.), International education in practice: Dimensions for schools and international schools. Routledge.

Chamberlain, S. (2010). AQA: Public perceptions of reliability. In Ofqual's Reliability compendium (Chapter 18). The Office of Qualifications and Examinations Regulation.

Coe, R., Searle, J., Barmby, P., Jones, K., & Higgins, S. (2008). Relative difficulty of examinations in different subjects. Report for SCORE (Science Community Supporting Education). CEM Centre, Durham University.

Cresswell, M. J. (1986). Examination grades: How many should there be? British Educational Research Journal, 12(1), 37-54.

Cresswell, M. J. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches. In H. Goldstein & T. Lewis (Eds.), Assessment: Problems, developments and statistical issues (pp. 57-84). Wiley.

Cresswell, M. J. (2000). The role of public examinations in defining and monitoring standards. Proceedings of the British Academy, 102, 69-120.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. Assessment in Education: Principles, Policy & Practice, 3(3), 265-286. https://doi.org/10.1080/0969594960030302

Dolan, R. P., Burling, K., Harms, M., Strain-Seymour, E., Way, W., & Rose, D. (2013). A universal design for learning-based framework for designing accessible technology-enhanced assessments (Research Report). Pearson. https://eric.ed.gov/?id=ED576691

Frith, D. S., & Macintosh, H. G. (1984). A teacher's guide to assessment. Stanley Thornes.

Gibbs, G. (1992). Teaching more students: Assessing more students. The Oxford Centre for Staff Learning and Development, Oxford Brookes University.

Gipps, C., & Murphy, P. (1994). A fair test? Assessment, achievement and equity. Open University Press.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 18(8), 519-521. https://doi.org/10.1037/h0049294

Goldstein, H. (1996). Group differences and bias in assessment. In H. Goldstein, & T. Lewis (Eds.), Assessment: Problems, developments and statistical issues (pp. 85-93). John Wiley & Sons Ltd.

Good, F. J., & Cresswell, M. J. (1988). Grading the GCSE. Secondary Examinations Council.

He, Q., Opposs, D., & Boyle, A. (2010). A quantitative investigation into public perceptions of reliability in examination results in England. In Ofqual's Reliability compendium (Chapter 19, p. 68). The Office of Qualifications and Examinations Regulation.

Hieronymus, A. N., & Hoover, H. D. (1986). Iowa tests of basic skills: Manual for school administrators. Riverside Publishing Company.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1983). Effects of achievement expectations and handwriting quality on scoring essays. Journal of Educational Measurement, 20(1), 65-70. https://www.jstor.org/stable/1434941

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. Journal of Applied Psychology, 71(2), 327-333. https://doi.org/10.1037/0021-9010.71.2.327

国際バカロレア. (2019). 『国際バカロレア（IB）の教育とは』 International Baccalaureate Organization.

国際バカロレア. (2024). 「「IBの使命」」 International Baccalaureate Organization.

Lambert, D., & Lines, D. (2000). Understanding assessment: Purposes, perceptions, practice. Routledge Falmer.

Linn, M. C. (1992). Gender differences in educational achievement. In J. Pfleiderer (Ed.), Sex equity in educational opportunity, achievement, and testing. Educational Testing Service.

Llewellyn, D. (2014). Inquire within: Implementing inquiry- and argument-based science standards in grades 3–8 (3rd ed.). Corwin.

Meyer, A., Rose, D. H., & Gordon, D. (2014). Universal design for learning: Theory and practice. CAST Professional Publishing.

Mitra, S. (2011, October 13-16). Responsibility, leadership and education [Keynote address]. Heads of Schools Conference, Singapore.

Murphy, P. (1999). Learners, learning & assessment. Paul Chapman Publishing in association with The Open University.

National Research Council. (2011). Assessing 21st century skills: Summary of a workshop. The National Academies Press.

Newton, P. (2011). A level pass rates and the enduring myth of norm-referencing. Research Matters. https://doi.org/10.17863/CAM.100441

Newton, P. E. (2007). Clarifying the purposes of educational assessment. Assessment in Education: Principles, Policy & Practice, 14(2), 149-170. https://doi.org/10.1080/09695940701478321

Newton, P. E. (2012). We need to talk about validity [Paper presentation]. The National Council for Measurement in Education Annual Meeting, Vancouver, Canada.

Newton, P. E., & Shaw, S. D. (2014). Validity in educational and psychological assessment. London: SAGE Publications.

Peterson, A. D. C. (1971). New techniques for assessment of pupils' work. Council of Europe.

Peterson, A. D. C. (2003). Schools across frontiers: The story of the International Baccalaureate and the United World Colleges (2nd ed.). Open Court Publishing Company.

Popham, W. J. (1978). Criterion-referenced measurement. Prentice Hall.

RAND Corporation. (2012, April). Teaching and learning 21st century skills: Lessons from the learning sciences. Asia Society. Retrieved January 1, 2017, from http://www.rand.org/pubs/external_publications/EP51105.html

Rao, K., Currie-Rubin, R., & Logli, C. (2016). UDL and inclusive practices in IB schools worldwide. CAST Professional Learning.

Schleicher, A. (2016). The case for 21st-century learning. OECD. Retrieved April 4, 2025, from https://web-archive.oecd.org/2012-06-14/61660-thecasefor21st-centurylearning.htm

Shepard, L. A. (1992). Commentary: What policy makers who mandate tests should know about the new psychology of intellectual ability and learning. In B. R. Gifford, & M. C. O'Connor (Eds.), Changing assessments: Alternative views of aptitude, achievement and instruction (pp. 301-328). Kluwer Academic Publishers.

Sireci, S. G. (2007). On validity theory and test validation. Educational Researcher, 36(8), 477-481. https://www.jstor.org/stable/4621103

Smith, D. J., & Tomlinson, S. (1989). The school effect: A study of multi-racial comprehensives. Policy Studies Institute.

Snyder, B. R. (1971). The hidden curriculum. MIT Press.

Surgenor, P. (2010). Teaching toolkit: Effect of assessment on learning. UCD Dublin.

Vygotsky, L. S. (1962). Thought and language. MIT Press.

Wiliam, D. (1993). Validity, dependability and reliability in National Curriculum assessment. The Curriculum Journal, 4(3), 335-350. https://doi.org/10.1080/0958517930040303

Winkley, J., & Cresswell, M. (2011). Introduction to the concept of reliability. In Ofqual's Reliability compendium (Chapter 1). The Office of Qualifications and Examinations Regulation.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. Journal of Child Psychology and Psychiatry, 17(2), 89-100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Wood, D. (1998). How children think and learn: The social contexts of cognitive development (2nd ed.). Blackwell Publishing.

Zanga, G., & De Gioannis, E. (2023). Discrimination in grading: A scoping review of studies on teachers' discrimination in school. Studies in Educational Evaluation, 78. https://doi.org/10.1016/j.stueduc.2023.101284

付録

用語解説

IB では、平易な言葉を使うように努め、専門用語の使用は極力避けるようにしています。この取り組みの一環として、本資料またはその他の評価関連資料で使われる主要な用語や概念の定義を用語解説にまとめました。用語の一部は、IB からのお知らせ等にも使われています。

用語	定義
学問的誠実性（Academic integrity）	個人の誠実さ、および指導、学習、評価の優れた実践を推進するための一定の価値観とスキル。
学問的不正行為（Academic misconduct）	特定の生徒あるいは他の生徒が、1 つまたは複数の評価要素において不公平な利益を得る結果となる、あるいはその可能性がある、故意または過失による行為。他の生徒の不利益となる行動もまた、学問的不正行為と見なされる。
到達度（Achievement level）	評価のために提出された成果物が、対応するレベルの説明と一致する場合に付与されるレベル。到達度は、評価規準の左欄に表示されている。
到達度のレベルの説明（Achievement level descriptors）	評価のために提出された成果物に対して、到達度ごとに期待される特徴を説明したもの。
管理者コンソール（Admin Console）	MYP のコンピューターを用いた試験で使われる安全なウェブサイト。MYP コーディネーターや指名された主任試験監督は、このウェブサイトから、デバイスの登録、試験パッケージへのアクセスといった重要な管理活動を行うことができる。管理者コンソールは、IB インフォメーションシステム（IBIS）内のリンクからアクセスできる。 注：DP と CP のデジタル試験については、デジタル試験システムを参照。
特別な事情（Adverse circumstances）	評価における生徒のパフォーマンスに悪影響を与える可能性があり、生徒のせいではない事情または事態。これには、重度のストレス、特段に困難な家庭の事情、忌引き、生徒の健康または安全を脅かし得る事象が含まれる。同一の事情が、複数の生徒または 1 つの学校の全生徒に影響する場合もある。以下については、特別な事情とは見なされない。

用語	定義
	・　生徒が登録されている学校側の過失 ・　受験上の配慮が承認され、試験において実施されたにもかかわらず、成績が改善しなかった場合
集約（Aggregation）	評点とスコアを合算して、最終的な結果を導き出すプロセス。
整合性（Alignment）	共有される学習の価値と学習への期待（指導計画）、教師の指導の仕方（授業方法）、および生徒が学ぶこと（評価計画）に関する、原則と実践における合意。
分析的マークスキーム（Analytic markscheme）	正しい解答、および評点を与えるべき場所が示されているマークスキーム。
評価（Assessment）	生徒が自らの能力を示すために完成させた課題。これらの課題は、学校ごとに作成され実施されるものと、IB に提出されるものがあり、指導と学習に関して判断を下すためのエビデンスを集めたものとみなされる。
評価要素（Assessment component）	1 つまたは複数の課題で構成され、全体的な評価（試験、作品ポートフォリオ、プロジェクト、研究課題など）の一部を成す。
評価内容開発者（Assessment content developer）	試験問題案、および試験に使用されるマークスキームの作成、精査、見直しを行う専門家。評価内容開発者はほとんどの場合上級試験官が努めるが、試験官とは別のスキルが求められる。
評価規準（Assessment criteria）	生徒のパフォーマンスを測定するための規準。
評価のサイクル（Assessment cycle）	試験をはじめとする評価の作成、実施、採点における手順。IB は、各試験セッションからの学びを次のセッションの改善に活かすため、これは循環型のプロセスと見なされる。
評価への応答（Assessment response）	評価課題に対する反応として生徒が作成したものすべて。
評価のストラテジー（Assessment strategy）	生徒の学習に関する情報を集める際に使用される方法またはアプローチ。例えば観察、オープンエンド型の課題、生徒の解答など。
評価課題（Assessment task）	生徒が評価を受けるために取り組む活動、または一連の活動。

用語	定義
評価ツール（Assessment tool）	生徒のパフォーマンスや理解に関する情報を集めるための方法。
到達度準拠（Attainment-referencing）（または軽度の目標基準準拠（Weak criterion-referencing））	生徒の到達度を、事前定義された到達度の説明（規準）および過去の受験者群のパフォーマンスと比較すること。基準の維持に関して、IB が採用しているアプローチ。
異例な解答（Atypical response）	課題に対して通常提出される解答とはまったく異なる解答。異例な解答の例には、未完成の成果物、指示に従っていない成果物、予期されない解答、問題の多い成果物、不正行為などが含まれる。
Authentication（生徒本人が取り組んだものであることの認証）	成果物が生徒によって完成されたことを実証するエビデンスおよびプロセス。例として、生徒の成果物が本人によるものであることを確認する教師と生徒の署名などが含まれる。
自動採点（Automarking）	テクノロジーを使用して、事前に定義されたマークスキームに照らして生徒の課題を評価するプロセス。ここでは、解答が客観的に正しいか間違っているかを判断する。
逆流効果（Backwash effect）	プロセスの後工程が前工程の実施に及ぼす影響のこと。教育の文脈においては、通常、学習と指導が評価によって影響を受けることを指す。
帯域チェッカー（Bandwidth checker）	DP と CP のデジタル試験に関して、学校のインターネットの帯域が複数の生徒の同時受験に対応できることを確認するために DP 校および CP 校の管理者が使用するオンラインツール。試験セッションのために設定を最適化できるよう、時間的余裕をもって帯域チェッカーを実行することが推奨される。
バイアス（bias）	ある設問または課題に関して、評価の対象であるスキルや知識における能力の差以外の理由で特定の集団が平均とは異なるパフォーマンスを見せる要因。
私用デバイスの持ち込み（Bring your own device）	生徒が自分のデバイスを使って学校の課題に取り組むこと。学校での使用に適したデバイスの購入について、保護者向けの手引きを学校が提供する。
試験官長（Chief examiner）	最上位の試験官。長期にわたって、かつ教科内の学習分野の間で、同じ基準が維持されていることを確認する責任を負う。

用語	定義
主任試験監督者（Chief invigilator）	DP および CP：主任試験監督者は学校におけるデジタル試験を監督し、デジタル試験システムでの管理業務、および試験前と試験中の技術的な問題に対応する責任を負う。主任試験監督者は学校の教職員でなければならず、プログラムコーディネーターが My IB においてこの役割を割り当てる。 プログラムコーディネーターとは別の役割であるが、プログラムコーディネーターと密接に協力し、試験プロセスの円滑な実施を支える。場合によっては、プログラムコーディネーターが主任試験監督者を兼任することもできる。 MYP：主任試験監督者の任命は任意。プログラムコーディネーターは、コンピューターを用いた試験に関する管理業務の補佐役を任命することができる。
指示用語（Command term）	設問の中で使われている、確認対象となる評価目標を説明する言葉。
同等性（Comparability）	具体的な結果がどこまで他の結果と同等と見なされ得るかの度合い。通常、年度ごと、または科目ごとの比較で使われる。
互換性チェッカー（Compatibility checker）	デバイスがデジタル試験の技術的要件を満たしていることを確認するために、MYP のコンピューターを用いた試験、および DP と CP のデジタル試験で使われる独立型アプリケーション。評価に使われるすべてのデバイスで、このアプリケーションを実行する必要がある。互換性チェックに合格していないデバイスを試験に使用することはできない。
構成の関連性（Construct relevance）	意図したスキルや知識を評価が実際にどこまで試すかの度合い。
コース（Course）	あらかじめ決められた学習期間内に行う一定数の授業、講義、指導時間など。学校は、科目特有のコースと学際的なコースを通じて、科目の学習と指導を構成する。
科目別履修証（Course results）	IB の評価結果を表す主要な文書。生徒が履修した科目と取得した成績（1〜7、もしくは A〜E）が記載される。また、「コア」要素において取得した成績など、その他の重要な評価結果も記載される。 さらに、生徒の氏名、生徒コード、セッション番号、成績が授与されたセッション、履修証の発行日、生徒を登録した学校名（転校があった場合は転校先）も記録されている。

用語	定義
目標基準準拠（Criterion-referencing）（または到達度準拠（Attainment-referencing））	成績を付与するために、生徒の到達度を事前定義された到達度の説明（規準）と比較すること。
評価規準に準拠した絶対評価（Criterion-related assessment）	あらかじめ合意された評価規準に照らして到達度を決定する評価プロセス。基準は固定され、受験者全体の到達度によって変化することはない。
確定済みの評点（Definitive mark）	主任試験官が、評価のために提出された生徒の成果物の特定の部分に付与した評点。他のすべての試験官はこの採点を模倣することを目標とする（**品質モデル**も参照）。
デバイス（Device）	IB のデジタル評価に使用されるノートパソコンまたはデスクトップコンピューター。 IB では、タブレットなどのプラットフォームとの互換性が現在継続的に検討されていることを踏まえ、「コンピューター」ではなく「デバイス」という用語を使用している。
（評価における）差異の明確化（Differentiation（assessment））	異なる能力レベルを示している生徒を区別すること。
（学習と指導における）個別最適化（Differentiation（learning and teaching））	多様な学習者のニーズに応えるために、内容、プロセス、成果物の調整を通して指導ストラテジーを変更すること。
デジタル試験の解答（Digital examination responses）	IB のデジタル評価では、生徒が完成させた成果物の呼称はプログラムによって変わる。 MYP では「生徒の解答ファイル」を呼ばれる。 DP と CP では「生徒の解答」と呼ばれる。 どちらも、試験中に作成され評価のために提出された最終版のデジタル出力を指す。紙ベースの試験では「答案（スクリプト）」と呼ばれる。

用語	定義
デジタル試験システム（Digital Examination System）	DP と CP のデジタル試験を実施するために学校が使用するオンラインシステム。
デジタル試験（Digital examinations）	MYP では、コンピューターを用いた試験が 10 年以上前から実施されている。デジタル仕様に特化して設計された、さまざまなメディアを使ったインタラクティブな課題であり、セキュリティーが確保された試験環境において実施される。 DP および CP におけるデジタル試験は正式な最終年次の評価であり、生徒はセキュリティーが確保された試験環境においてデバイスを使用して受験する。従来の紙ベースの試験に替わるものとして、段階的に導入される予定である。
学問分野（Discipline）	学習分野や学術研究の領域。指導の目的によって知識を整理する方法（一般には、MYP と DP における評価という実践的な目的に基づき、「科目」として知られる）。
ダイナミックサンプリング（Dynamic sampling）	モデレーションに際して品質確認を効果的に活用するためのプロセス。
e アセスメント（eAssessment）	コンピューターを用いた評価、e ポートフォリオ、「パーソナルプロジェクト」という 3 つの要素から成る MYP の評価。
e マーキング（eMarking）	試験官が試験の解答をデバイスの画面上で直接採点するプロセス。
成績照会サービス（Enquiry upon results）	学校の要請を受けて行われる評点の確認。
e ポートフォリオ（ePortfolios）	MYP の生徒の内部評価の試験やコースワークを学校がアップロードして、IB による外部モデレーションを受けるシステムまたはプロセス。
試験（Examination）	試験は評価の一形態であり、さまざまな種類の課題（短答形式、論述形式、問題解決形式、分析形式の問題、実習や口述課題など）が含まれる。生徒は、厳しく管理された状況において所定の時間内に解答しなければならない。
試験監督者（Examination invigilator）	試験環境を監督し、試験の安全性確保に寄与する担当者。

用語	定義
試験問題（Examination paper）	試験において生徒が解答しなければならない課題群および設問群。コンピューターデバイスを用いた試験を指す場合もあれば、紙と鉛筆を使った試験を指す場合もある。
試験セッション（Examination session）	試験が実施・採点される期間。IB では年 2 回、5 月と 11 月に試験セッションを実施している。
試験官（Examiner）	生徒の外部評価に評点をつける担当者。
試験官の再採点（Examiner re‐mark）	試験官のつけた評点が、一貫性を欠いていたり、必須の基準から大きく外れていたりすることが分かった場合に、生徒の解答（評価の構成によって、解答全体のこともあれば設問項目グループのこともある）の採点をやり直すプロセス。多くの場合、モデレーションの失敗を受けて行われる。
例外的な事情（Exceptional circumstances）	受験上の配慮要件がある他の生徒にはあてはまらない、一般的ではない状況。どのような状況が「例外的」な事情に該当し、ゆえに一定の受験上の配慮を適用する根拠となるかの判断は、IB の裁量に委ねられる。
外部評価（External assessment）	生徒を担当する教師ではなく、IB が設定し採点する評価。
外部モデレーション（評価の適正化）（External moderation）	モデレーションを参照。
外部評価された（Externally assessed）	IB が完全に評価を行った成果物。
練習（Familiarization）	IB のデジタル試験環境とツールの使い方に慣れるために、コンテンツフリーの練習用システムを使って操作方法を学ぶプロセス。学校は、デジタル試験に向けて生徒が練習ツールに定期的にアクセスできるようにする責任を負う。 練習環境は MYP と DP および CP において用意されており、いずれもプログラム・リソース・センターから入手できる。
練習ツール（Familiarization tool）	デジタル試験環境を模倣したコンテンツフリーのシミュレーション。科目固有の内容に触れることなく、生徒がツールの使い方や試験のレイアウトを理解し、プラットフォームの操作方法を練習できる。

用語	定義
最終評価（Final assessment）	コース修了時に行われる、成果物に対する総括的評価。
形成的評価（Formative assessment）	指導を導き生徒のパフォーマンスを向上させるために情報提供することを目的とした継続的な評価。
成績（Grade）	生徒の到達度を説明するもの。評価のために提出された成果物に対する最終成績は、1（最低）から 7（最高）までの範囲で与えられる。成績とは、生徒が見せた全体的な質に対する IB の判断を示すもので、複数の年度や科目を通じて一貫性を有している。成績は、生徒のパフォーマンスの質を示すものであり、試験、学年、科目を問わず同じ意味をもつ必要がある。評価要素のセッション間の相対的難易度や基準の変更などを考慮に入れて決定される。
成績付与（Grade award）	評点を成績に換算する方法を決定するプロセス。これにより、受験したセッションにかかわらず、成績の意味を同じにすることができる。
成績区分（Grade boundary）	ある成績と次の成績の間の境界線を示す。たいていは、特定の成績に相当する評価規準の評点合計の最低値と最高値を示すために用いられる。
成績評価の説明（Grade descriptors）	各成績で生徒が達成すべき質を明確化したもの。科目や教科に特有のものもあれば、プログラム全体に共通するものもある。いずれの場合でも同じ特徴を説明していなければならない。具体的な例は、科目に特有の文脈で成績評価の説明が何を意味するのかを示すためだけに使われる。
包括的な規準（Holistic criteria）	生徒の成果物を評価する際に、個別の要素（例：コミュニケーション、科目の知識、議論の質）を一つひとつ検討するのではなく、成果物全体を 1 つの結果と見なすアプローチ。
IB インフォメーションシステム（IB information system）	パスワードで保護されたウェブサーバーを経由して、プログラムコーディネーターが管理手続きを行い、IB からニュースや情報を取得するためのシステム。
インクルーシブな評価（Inclusive access）	すべての生徒が当該科目における能力を公正な方法で実証できるよう、各生徒のニーズを考慮した評価。

用語	定義
受験上の配慮（Inclusive access arrangements）	評価のプロセスにおいて、配慮を必要とする生徒のために試験の実施条件などを変更・追加すること。これにより、生徒が自分の到達度をより公正な方法で実証できるようになる。
学際的な評価（Interdisciplinary assessment）	1 つの評価で 2 つ以上の学習分野または学問領域を組み合わせること、またはそれらが関わっていること。DP では、学際的な科目は 1 つの科目を通して 2 つの教科の要件を満たすものを指す。MYP では、学際的な学習は、教科内、教科間のいずれでも展開することができる。MYP の学際的な外部評価は、必ず複数の教科を対象とする。
内部評価（Internal assessment）	教師が学校で行う評価。
内部評価の標準化（Internal standardization）	学校内で同じ科目を担当するすべての教師が同じ基準で評価を行うようにするためのプロセス。
内部評価された（Internally assessed）	生徒を担当する教師が評価した成果物。内部評価された成果物のうち、サンプルとして IB に選ばれたものがモデレーションを受ける。
成績交付（Issue of results）	評価結果に基づく成績を IB から生徒に付与するプロセス。
項目（Item）	1 つの評価における適切な最小単位。個別でありながら全体性をもつ、評価の対象となる部分。各科目の IB 資料『指導の手引き』の「シラバスの概要」または「外部評価の詳細」に基づいて説明と周知が行われ、評価に含まれることが期待されている。
判断（Judgement）	生徒の成果物を個々の評価規準に基づいて検討した結果。
ロックダウンモード（Lockdown mode）	デジタル試験中に関係のないアプリケーション、ウェブサイト、ファイルにアクセスできないようにする、セキュリティーが確保された試験設定。 MYP：ロックダウン機能は試験パッケージに内蔵され、評価の開始と同時に自動で起動する。 DP と CP：ロックダウンモードは、Windows と Mac では Safe Exam Browser を使うことで実施される。学校が管理する Chromebook については、Chrome Kiosk モードで実行される。生徒のデバイスの設定管理は学校の責任である。

用語	定義
不正または過失（Maladministration）	IB ワールドスクールによる、IB の規則に違反する行為、および IB 試験および評価の正当性を脅かすおそれのある行為。これは、評価への取り組み前、取り組み中、取り組み後、または試験の実施前、実施中、実施後に起こる可能性がある。
違反行為（Malpractice）	学問的誠実性の原則に違反するあらゆる行為（剽窃など）。
管理のしやすさ（Manageability）	評価および個々の課題が生徒または学校に課す負担の度合い。管理のしやすさの例として、評価の長さ、評価を実施するための設備や資材、資格に必要な評価の数などが挙げられる。
評点（Mark）	特定の課題に対する生徒の解答の質を反映した数値。評点は、設問に対して生徒が正しく解答できた割合を表すために、規準、マークバンド、マークスキームに従って付与される。配点は、設問および試験ごとに異なる。
マークバンド（採点基準表）（Markband）	所定の質を示した生徒の解答に対して付与されるべき評点の範囲を示したもの。
マークスキーム（採点基準）（Markscheme）	任意の成果物に対して規準のレベルを決定するための指針。
成績がつけられない場合の手順（Missing grade procedure）	MYP の生徒について、その成果物に基づき IB が正確かつ公正な成績を算出できない場合に、当該生徒に成績を付与するためのメカニズム。IB または（学校以外の）第三者の行動の結果として生徒の成績のエビデンスが不足し、もう一度評価を受けるよう要請することが合理的でない場合には、この手順をとることが適切となる。
評点がつけられない場合の手順（Missing mark procedure）	DP および CP の生徒について、その成果物に基づき IB が正確かつ公正な評点を算出できない場合に、当該生徒に評点を付与するためのメカニズム。IB または（学校以外の）第三者の行動の結果として生徒の評点のエビデンスが不足し、もう一度評価を受けるよう要請することが合理的でない場合には、この手順をとることが適切となる。
モデレーション（評価の適正化）（Moderation）	共通の評価基準が守られていることを確認するためのプロセス。評価のために提出された成果物のサンプルを確認し、必要に応じて評価者の評点を調整する。

用語	定義
モデレーション係数（Moderation factor）	教師がつけた評点（規準レベル）を共通の評価基準に合わせるために適用する、計算上の調整。
モデレーションサンプル（Moderation sample）	採点結果が必須の基準に沿っていることを確認するために IB に提出される成果物のサンプル。
調整済み試験問題（Modified paper）	特別なニーズのある生徒がそうしたニーズのない生徒と平等な立場で評価を受けられるようにするために評価に対して加える変更。例えば、フォントの種類や形を変更するといったことが含まれる。この種の変更によって問題の性質が変わることがあってはならない。
多肢選択問題（Multiple-choice question）	与えられた複数の選択肢から、正しい答えを選ぶことが求められる問題。
MYP のコンピューターを用いた試験（MYP on-screen examination）	規定時間内に実施される、さまざまなメディアを用いた MYP 専用の正式なデジタル評価。安全性が確保されたデジタル環境で行われ、科目に特化した課題に生徒がデバイス上で対応する。 インターネット接続は必須ではないものの、オンライン上で実施することが強く推奨される。これにより、生徒の解答が自動で保存され、管理業務の負担が軽減される。多くの学校が、この理由によってオンライン上で試験を実施している。
集団基準準拠（Norm-referencing）	生徒のパフォーマンスを、評価の対象とされる生徒全体のパフォーマンスと比較（準拠）することで生徒の到達度を決定する方法。
目標（Objective）	評価対象となるスキル、知識、理解を説明した一連の記述の 1 つ。
試験問題（Paper）	**試験問題**を参照。
パイロット科目（Pilot subject）	評価中の科目。評価結果が良ければ、全体的に導入される。
剽窃（Plagiarism）	故意であるかどうかにかかわらず、適切、明確、かつ明示的な認知をせずに他人の考え、言葉、成果物を自分のものとして提示すること。
練習用スクリプト（答案）（Practice script）	評価のために提出された成果物のうち、標準化プロセスにおいて選定・採点されたもの。採点の際に守るべき基準を示すために試験官に提供される。

用語	定義
予測可能性（Predictability）	将来何が起こるか、また、それがいつ起こるかを予測できる度合い。評価の文脈においては、試験問題でどのような問いが出題されるか、また、それがいつ出題されるかを学校が予測できることを指す。
主任試験官（Principal examiner）	ある要素の評価を主導する責任者。主任試験官は評価の採点基準を設定する役割を果たし、通常、評価の作成者も兼任する。 MYP で主任試験官が果たす役割は他の試験システムとはやや異なり、特定の学習分野の責任者として、評価を設計するチームを統率し、基準の設定と維持の責任を負い、さらに試験官チームリーダーのメンターを務める。
認定用スクリプト（答案）（Qualification script）	実際の採点に入る前に試験官が採点の必須基準を理解していることを確認する目的で、主任試験官によって指定された生徒の成果物の例。
品質モデル（Quality model）	生徒に正しい評価結果が付与されていることを確認するために IB が使用するアプローチ。主任試験官が問題ごとに正しい基準を設定し、すべての試験官がこの基準を踏襲する。外部評価においては、標準化を通じて試験官に見本を示し、認定用スクリプトを使って試験官が基準を理解していることを確認し、さらにシードスクリプトを使って定期的に試験官の採点をモニタリングすることで、この品質モデルが実践される。
設問（Question）	生徒が科目における能力を実証するために取り組む課題や活動。
問いバンク（Question bank）	一連の設問と、それに関連するトピックや予想される難易度についての情報をまとめたもの。問いバンクの情報を参考に、試験問題を作成することができる。IB では現在、問いバンクは使用していない。
設問項目グループ（Question item group）	同じ試験問題に含まれる 1 つまたは複数の関連する設問を、グループとして捉えたもの。試験官は、試験問題全体を採点するのではなく、特定の設問項目グループに特化して採点を行う。このアプローチをとることにより、生徒の成果物全体を 1 人の試験官が採点するよりも信頼性の高い評価結果が達成される。
信頼性（Reliability）	成果物が評価されるたびに、生徒が同じ結果を受け取る可能性。複数の試験官の間での信頼性を指すこともあれば、1 人の試験官の信頼性を指すこともある。

用語	定義
使用言語（Response language）	生徒が評価に解答する際に使う言語。
再試験（Retake）	1つまたは複数の試験を二度以上受験すること。評点を上げてIB MYP修了証やIBディプロマを取得すること、あるいはすでに授与された修了証の合計点を上げることを目的とする。
答案（スクリプト）（Script）	紙ベースの試験における生徒の解答。評価のために提出されたあらゆる成果物を指す場合もある。
シード（Seed）	主任試験官がすでに採点した生徒の成果物。試験官に割りあてられるスクリプトに無作為に追加される。試験官がシードで行った採点は、その試験官が一定の許容範囲内で基準を踏襲していることを確認するため、主任試験官の採点と比較される。ダイナミックサンプリングを使用するモデレーションでも、そのプロセスの一環としてシードが使われる。
上級試験官（Senior examiner）	主任試験官を補佐する経験豊富な試験官。
セッション（Session）	試験セッションを参照。
試験見本（Specimen examinations）	科目ごとのIB試験の完全な見本。試験見本には、実際の評価の内容が含まれ、最終評価の体裁およびレベルが再現されている。多くの場合、試験に近い状況で練習をするための教材として使用される。 MYP：試験見本パッケージがプログラム・リソース・センターに収載されている。 DPおよびCP：試験見本はプログラム・リソース・センター経由で入手可能となる。デジタル試験に関しては、実際のデジタル試験システム経由で追加のアクセスが付与される可能性がある。
基準（Standard）	特定のスコア、成績、結果に対応する、期待されるパフォーマンスのレベル。
標準化（Standardization）	モデレーターまたは試験官の間に共通の評価基準を浸透させるための協働のプロセス。
標準化会議（Standardization meeting）	採点に際しての必須基準を説明し、シードスクリプトを設定する目的で、主任試験官が主催する会議。

用語	定義
生徒（Student）	IB のコースまたは IB の教育プログラムに参加する人。
生徒の登録（Student registration）	プログラムコーディネーターが IB 評価のために生徒を登録するプロセス。
生徒の解答ファイル（Student response file）	MYP のコンピューターを用いた試験において、「生徒の解答」は生徒が試験中に完成させた成果物を指し、これにはマルチメディアが含まれる場合もある。解答を作成するプロセスも評価対象に含まれる。生徒の解答は電子ファイルとして提出される。これはしばしば「生徒の解答ファイル」を呼ばれる。
提出（Submission）	生徒が（または生徒に代わって学校が）最終成果物を IB に提出するプロセス。
総括的評価（Summative assessment）	生徒の能力や到達度を見極めるための評価。通常は、コースや単元の終了時に行われる。
システム要件（System requirements）	デバイスが IB のデジタル試験を安全かつ円滑に実行するために必要とされる最低限の技術要件。これらの要件は、評価中に発生したハードウェアの問題によって生徒が不利益を被ることがないように設けられている。 要件は、MYP と DP および CP を対象に公開されている。すべての生徒のデバイスが必須要件を満たしていることを確認するのは学校の責任である。
システム要件チェッカー（System requirements checker）	デバイスが DP および CP のデジタル試験の実行に必要な技術的要件を満たしていることを確認するために使用されるツール。学校は、時間的余裕をもってデバイスのテストを実施し、安全で円滑な評価体験が提供されることを確認するとよい。 このチェッカーは公開されているシステム要件を補完するものであり、試験当日ではなく試験の準備期間中に使われる。
IB 資料『教師用参考資料』（Teacher support material）	IB コースの要件に関する教師の理解を助ける付加的な情報。各科目の『指導の手引き』に示されている理論の理解と実施を助けるための実践的なアドバイスとサポートを提供することを目的としている。
チームリーダー（Team leader）	試験官のチームを率いる試験官。
テスト（Test）	「テスト」は一般的には試験を表す言葉として使われるが、学術論文や評価においては特別な意味をもつことがある。

用語	定義
	場合によって、生徒の評価全体が個別の要素に分けられ、それぞれ異なる時期に実施されることがある。これは「評価要素（コンポーネント）」と呼ばれる。評価要素の例として、口述試験、個別試験または内部評価が挙げられる。
許容差（Tolerance）	主任試験官の決定的な採点とは異なる、わずかなばらつき。IB は、試験官の採点がこの範囲に収まっていれば、正しい基準を踏襲していると見なす。採点は個人の判断に基づくものであり、経験豊富な試験官ですら、同じ成果物をあらためて採点すればわずかに異なる評点をつける可能性があることから、許容差を設けることは必要不可欠である。許容差は、評点の幅や問題の種類、さらには科目によって異なる。
評価のユニバーサルデザイン（Universal design for assessment）	一部の生徒に特別な対応をするのではなく、あらゆる生徒のニーズを理解して、その理解を基にすべての評価を開発すべきという考え方。学びのユニバーサルデザインに向けた IB の取り組みの一部を成す。
学びのユニバーサルデザイン（Universal design for learning）	すべての生徒に焦点をあてた、生徒の学びのための位置づけと枠組み。学習と関わり、学習に参画し、学習を表現するために複数の手段を提供することで、インクルーシブな環境を醸成することをねらいとしている。
妥当性（Validity）	評価や評価結果の使途が目的に合っているかどうかを説明する用語。
妥当性の議論（Validity argument）	評価の作成において下された意思決定に関して、評価が目的に適っていることを示すエビデンスと説明。
軽度の目標基準準拠（Weak criterion-referencing）（または到達度準拠（Attainment-referencing））	生徒の到達度を、事前定義された到達度の説明（規準）および過去の受験者群のパフォーマンスと比較すること。基準の維持に関して、IB が採用しているアプローチ。
IB の使用言語（Working languages）	IB が関係者に連絡する際に使用する言語。IB は、使用言語において、プログラムの実施に関する各種のサービスを提供することを約束している。IB の現在の使用言語は、英語、フランス語、スペイン語。

付録

印刷可能な資料

本資料全体を通して、印刷可能な資料が引用されています。目的の資料を見つけやすいよう、印刷可能な資料の一覧を以下に示します。

- 「評価サイクルにおける重要な役職の責任と義務」(PDF)
- 「ブルームの分類法」(PDF)
- 「倫理的な考え方の育成」(PDF)
- 「受験上の配慮を実施する」(PDF)
- 「内部評価の採点：モデレーションプロセスにおける教師への期待事項」(PDF)
- 「ダイナミックサンプリングを使用したモデレーション」(PDF)
- 「予測スコア：教師用ガイド」(PDF)
- 「品質モデルで必要とされる内部評価スクリプト（答案）の範囲」(PDF)
- 「総括的評価と形成的評価」(PDF)
- 「IB の評価の原則」(PDF)
- 「妥当性の鎖」(PDF)