和 2 年度　厚生労働科学研究費補助金（障害者政策総合研究事業）
分担研究報告書

療育手帳判定におけるアセスメント手法に関する心理測定学的検証

分担研究者　伊藤　大幸　中部大学　現代教育学部

研究要旨

本稿では、心理アセスメント検査の体系的な評価の枠組みである EFPA モデルについて概説した上で、療育手帳判定に用いられている 9 つの検査の心理測定学的性質について検証を行った。その結果、知能検査としてはウェクスラー式検査、適応行動尺度としては Vineland-II 適応行動尺度のみが、療育手帳判定の用途に適合した性能を有していることが示された。ビネー式検査や発達検査は、知的障害の診断基準にそぐわない比率指数（精神年齢と生活年齢の比）を用いていること、サンプリングの方法が適切でないこと、サンプルの代表性に関する情報が提示されていないこと、信頼性・妥当性について十分な根拠が示されていないことなどから、対象者の処遇を左右する high-stakes な判断である療育手帳判定に用いることは適切でないと結論づけられた。

また、成人の一般サンプルおよび知的障害者のデータに基づく検証により、知能、適応行動をそれぞれ単独で用いるよりも、両者の合成値を用いることが知的障害者の判別の精度を高めることが示された。こうした結果に基づけば、現在、多くの自治体で標準化された方法によって適応行動の評価がなされていないことは重大な問題であり、今後、適応行動尺度の普及を図っていくことが必要であると考えられる。

A.　研究目的

療育手帳制度は、1973 年の厚生事務次官通知「療育手帳制度について」に基づく制度であり、知的障害児者に対する主要な施策の一つであるが、制度開始以来、統一的な規定は設けられておらず個々の自治体が独自に規定・運用を行っている。精神医学領域では、1980 年代以降、DSM（APA, 2013）や ICD（WHO, 2015）などの操作的診断基準が整備されてきており、知的障害の診断と重症度の判定は、知能および適応行動に関する標準化検査の結果に基づいて行うことが求められている。こうした流れに一致して、各自治体の療育手帳判定においても標準化された知能検査や発達検査が用いられているが、複数の実態調査（柴田, 2004; 吉村ら, 2019）や本研究班の小林（2021）の分担報告が明らかにしているように、使用される検査や判定の基準は自治体によって異なっている。標準化された検査であっても、標準化の手続きがいつ行われたのか、測定の精度（信頼性・妥当性）がどのように検証されているのかは検査によって様々であり、必ずしも全ての検査が同等の心理測定学的な質を備えているとは限らない。また、知能検査と適応行動尺度という 2 つの独立したアセスメントの結果をどのように組み合わせて総合的な判定を行うかについても、自

治体によって運用のあり方が大きく異なっている。

こうした現状を踏まえ、本稿では、心理アセスメント検査の質を評価するための体系的な枠組みである欧州心理学者連合会（European Federation of Psychologists' Associations: EFPA）検査評価モデル（Evers, Hagemeister, & Hostmaelingen, 2013）について概観した上で、療育手帳判定に用いられている各種検査の心理測定学的性質を包括的に評価し、判定業務での使用に関する推奨の程度を判定する。加えて、知的障害のアセスメントに用いられる知能検査と適応行動尺度の結果をどのように組み合わせて総合的な判定を行うことが望ましいのかについて、統計学的観点から予備的な検証を行う。

B. 研究方法

第一に、心理アセスメント検査の体系的な評価の枠組みである EFPA 検査評価モデルについて概観する。第二に、療育手帳判定に用いられている各種検査の心理測定学的性質を EFPA 評価モデルに基づいて評価し、判定業務の使用における推奨の程度を判定する。

第三に、定型発達者と知的障害者の知能および適応行動のデータを使用して、知的障害の判定の方法について統計学的観点からの予備的検証を行う。このデータは、厚生労働省令和 2 年度社会福祉推進事業「日常生活支援住居施設の対象者選定のためのシステムに関する調査研究事業」（代表者：辻井正次）によって収集され、一般サンプル 418 名（男性 208 名、女性 211 名；平均年齢 49.8 歳）および知的障害者 33 名（男性 20 名、女性 13 名；平均年齢 49.1 歳）より得られたものである。知能および適応行動の測定には、同事業によって開発された ABIT（Adaptive Behavior and Intelligence Test）を用いた。ABIT は、タブレット端末を用いた包括的アセスメントツールであり、知能については対象者が 3 つの課題に取り組む形式、適応行動については対象者をよく知る同居者・介護者が 35 の項目に評定する形式で測定された。WAIS-IV および Vineland-II 適応行動尺度との相関や複数の臨床群の識別精度に基づいて妥当性が検証されている。本研究の手続きは、中京大学現代社会学部倫理審査委員会の審査と承認を得た。

C. 研究結果

1．EFPA 検査評価モデル

ヨーロッパ地域の 34 の心理学系学会の統括組織である EFPA は、心理アセスメント検査（質問紙尺度を含む）の評価の枠組みとして、EFPA 検査評価モデル（EFPA Test Review Model）を示している（EFPA, 2013）。このモデルは、検査のユーザーがアセスメントに関して正しい意思決定を行うことに加え、専門家が検査の改善を図ることを支援するものである。EFPA 検査評価モデルの初版は、各国（イギリス、スペイン、オランダ）の心理学会における検査評価ガイドラインをもとに 2002 年に編集された（Bartram, 2002a, 2002b）。その後、数回の改訂を経て、2013 年に現行版（EFPA, 2013）が発表された。

EFPA 検査評価モデルは、「原理の説明および提供される情報の質」、「検査用具の質」、「基準値（norms）」、「信頼性」、「妥当性」、「コンピュータによるレポートの質」という 6 つのセクションから構成され、それぞれについて多数の評価項目が設定されている。また、個々の項目の個別評定に基づいて

セクションごとの総合評定が行われる。各セクションの総合評定で 0 や 1 の評定が与えられた場合、実践（臨床、教育など）における検査の使用は推奨されないとしている。

本稿では EFPA 検査評価モデルの 6 つのセクションのうち、心理測定学的に特に重要な意味を持つ「基準値」、「信頼性」、「妥当性」の 3 つのセクションについて評価を行う。各セクションの評価項目と評定基準を付録に示した。各項目の評定は原則的に表 1 に示す評定システムによって行われるが、一部項目では 0〜4 の評定値が設定されておらず、個々の検査の用途や開発の文脈に応じて弾力的に評価を行う形式を取っている。

1）基準値

基準値（norms）とは、アセスメントによって得られた粗点に意味づけを与えるための基準となる値であり、大きく 2 つの種類に分類される。一つは、準拠集団における粗点の分布に基づいて得られる基準値であり、これに基づく評価は集団基準準拠評価（norm-referenced interpretation）と呼ばれる。もう一つは、単一または複数の数値として与えられる基準値である。これはさらに、特定の領域の能力やスキルが習得されたか否かを評価する領域準拠評価（domain-referenced interpretation）と、実証研究により得られたカット得点に基づいて評価を行う数値基準準拠評価（criterion-referenced interpretation）に分けられる。EFPA 評価モデルでは、これら 3 種類の評価の基準値について別個に評価項目を設定しているが、知的障害の判定に用いられる知能検査や適応行動検査は基本的に集団基準準拠評価を採用しているため、ここでは集団基準準拠評価に関する評価項目にのみ焦点を当てる。

集団基準準拠評価の基準値については 9 つの評価項目（9.1.1〜9.1.9）が設定されている。9.1.1 では、利用地域に適合したサンプルに基づく基準値であるか否かが評価される。利用地域において、母集団の特徴に合致したサンプルが得られているほど、高い評定値が得られる。9.1.2 では、年齢区分ごと、性別ごと、民族ごとの基準値など、利用場面に適した基準値が示されているか否かを評価する。

9.1.3 と 9.1.4 では基準値を得たサンプルのサイズが評価される。基準値の構築には大きく 2 つの方法がある。古典的な方法では、個々のサンプルの得点分布にのみ基づいて基準値を求める。一方、近年よく用いられる連続的基準化と呼ばれる方法では、サンプルを複数の年齢（月齢）区分の下位サンプルに分け、下位サンプルごとに分布の特徴を表す何らかのパラメータ（平均値、標準

表 1 EFPA 検査評価モデルの評定システム

評定	説明
[n/a]	この検査には適用できない観点である
0	十分な情報が提供されていないため評定できない
1	不適格
2	適格
3	よい
4	優れている

偏差、尖度、歪度など）を求めた後、下位サンプル間でパラメータ推定値の平滑化処理を施すなどした上で、年齢区分ごとの基準値を得る。平滑化処理により隣接するサンプルの情報も利用できるため、下位サンプルごとのサイズは古典的方法よりも少なく抑えられる。古典的基準化では 1000 以上、連続的基準化では（8 つの下位サンプルに分けた場合）下位サンプルにつき 150 以上のサイズであれば最高評価となる。

9.1.5 ではサンプル抽出の手続きが評価される。ここでは 0〜4 の評定値は設定されておらず、どのような方法が用いられたかが記録され、後の項目の評定や総合評定の判断材料として用いられる。サンプリングの方法は大きく確率的サンプリング（母集団を構成する各成員が調査対象となるか否かの確率が属性によって変わらない方法）と非確率的サンプリングに分けられ、統計的観点からは前者の方法がより望ましいが、現実的な制約により後者の方法が用いられることも多い。確率的サンプリングは、純粋にランダムにサンプリングを行うランダムサンプリング、母集団ごとにサイズを決定する系統的サンプリング、特定の変数で層化を行う層化サンプリング、集団単位でサンプリングを行うクラスターサンプリング、クラスターごとにランダムサンプリングを行う多相サンプリングに分けられる。非確率的サンプリングは、単純に検査された全ての個人をサンプルに含める簡便サンプリング、あらかじめ属性ごとに必要数を割り当てる割当サンプリング、参加者に次の参加者を集めてもらうスノーボールサンプリング、特定の集団に参加を依頼する目的的サンプリングに分類される。

9.1.6 では基準値サンプルの代表性について評価を行う。データがランダムサンプリングによって得られている、サンプルと母集団の基本属性の構成が十分に記述されている、それらの属性に関して高い代表性が認められるなどの要件が揃うほど、高い評定が与えられる。

9.1.7 では、マイノリティ集団での得点差や性別、年齢の効果について、使用と解釈に必要な情報が十分に提供されているか否かが評価される。

9.1.8 では標準化研究が行われた時期が評価され、10 年以内であれば最高評定が与えられるが、20 年以上前であると不適格の評定となる。

9.1.9 はパフォーマンス検査（知能検査など課題へのパフォーマンスを評価する検査）にのみ適用される項目で、練習効果（複数回の検査を実施したときに得点が上昇する効果）について十分な情報が与えられているか否かを評価する。

これらの各項目の評定に基づいて、基準値の総合的な適格性が評価される（9.3）。集団基準準拠評価の総合的評定は、サンプルサイズに関する評定（9.1.3 または 9.1.4）を超えてはならないと定められている。サンプルサイズ以外の情報として特に重要視されるのは、サンプルの抽出方法（9.1.5）や代表性（9.1.6）と標準化研究の実施時期（9.1.8）である。

2）信頼性

信頼性は、測定値に占めるランダムな測定誤差の分散の小ささを表す。信頼性に関する基準は、集団に関する意思決定を目的とする場合と個人のアセスメントを行う場合で異なり、前者より後者で厳しいものとなる。信頼性の評価の方法には、内的整合性、検査―再検査信頼性、等価信頼性、項目反応理論による方法、評定者間信頼性など

の方法があり、検査の性質や用途に応じて複数の方法を併用して評価を行うことが一般的である。

10.1 では、信頼性に関して提供されたデータについて記録する。評定値の設定はなく、信頼性係数や測定標準誤差の報告があるか、また、複数の種類の指標が報告されているかが記録される。

10.2 では、内的整合性に基づく信頼性について 4 つの項目により評価を行う。内的整合性は、尺度を構成する項目間の相関に基づいて信頼性を推定する方法であり、項目間の相関が高いほど、また、尺度に含まれる項目数が多いほど、係数が高くなる性質がある。10.2.1 では、内的整合性の推定に用いられたサンプルのサイズについて評価する。複数の大規模 (200 以上) な研究により推定が行われている場合に最高評定となる。10.2.2 では、報告された信頼性係数の種類が記録される。広く用いられている α 係数の他、因子分析に基づく ω 係数などの種類がある。10.2.3 では、信頼性係数の値が評価される。.70 以上であれば「適格」、.80 以上であれば「よい」、.90 以上であれば「優れている」と評定される。10.2.4 では、係数の推定に用いられたサンプルの特徴について記録する。実際の利用対象に一致したサンプルであるか、適合しない場合、推定にどのようなバイアスが生じているかを判定する。

10.3 では、検査—再検査信頼性について評価を行う。検査—再検査信頼性は、一定期間 (通常は数週間程度) を挟んで同一の検査を複数回実施し、測定値の安定性に基づいて検査の信頼性を評価する方法である。評価項目は内的整合性とほぼ同様であるが、係数の種類に関する項目はなく、代わりに検査—再検査の間隔に関する項目 (10.3.3) が設定されている。また、係数の値について、内的整合性よりも基準が低く設定され、.60 以上で「適格」、.70 以上で「よい」、.80 以上で「優れている」と評定される。

10.4 では、等価信頼性について評価を行う。等価信頼性は、平行検査と呼ばれる複数の項目セットが設定されている場合に、それらの測定値の相関に基づいて信頼性の評価を行う方法である。評価項目は内的整合性と同様であるが、係数の種類に関する項目の代わりに、複数のバージョン間の平行性の仮定について評価する項目 (10.4.2) が設定されている。複数のバージョン間で、平均値、分散、相互の相関が一致しているほど高い評定が与えられる。

10.5 では、項目反応理論に基づく信頼性が評価される。項目反応理論は、検査項目への応答という顕在変数に基づいて、被検者の特性や項目の難易度、識別力などの潜在的なパラメータを推定するための理論である。評価項目は内的整合性と同様であるが、サンプルの特徴に関する項目はなく、係数の種類やサイズについても内容が異なる。

10.6 では、評定者間信頼性について評価が行われる。評定者間信頼性は、検査が何らかの判断のプロセスを含む場合に、評定者間の測定値の一致に基づいて信頼性の評価を行う方法である。評価項目は内的整合性と同様であるが、サンプルの特徴に関する項目はなく、係数の種類についても内容が異なる。また、係数のサイズについては検査—再検査信頼性と同一の基準が設定されている。

10.7 では、その他の方法に基づく信頼性について評価を行う。古典的な折半法に基づく推定などが含まれる。サンプルサイズ、用いられた手法、得られた結果について、記録および評価が行われる。

10.8 では、10.1 から 10.7 の記録と評価に基づいて、検査の信頼性を総合的に評価する。単純な平均値を取るのではなく、検査の性質や用途と照らし合わせて弾力的な判断を行う。具体的には、検査が個人のアセスメントに用いられるか集団に関する意思決定に用いられるか、決定の性質は high-stakes か low-stakes か、複数の方法に基づく信頼性が報告されているか、信頼性係数だけでなく測定標準誤差も提供されているか、手続き的な問題（サンプルサイズ、研究の数、評定者の数、検査―再検査の間隔など）がないか、わかりやすい報告がなされているか、などの観点を考慮して総合的に判断する。

3）妥当性

妥当性は、検査の系統的な誤差（バイアス）の小ささを表す。EFPA 評価モデルでは、構成概念妥当性と基準関連妥当性という２つのセクションで妥当性が評価される。構成概念妥当性は検査の測定値が測定を意図した概念をどの程度忠実に反映しているか、基準関連妥当性は検査の測定値が実在の外在基準（他の尺度ではない）とどの程度関連するか（例えば入社試験の測定値が職業上の成功をどの程度予測するか）を意味し、近年では、前者が後者を包含するという考え方が一般的である。知的障害の判定に用いられる知能検査や適応行動検査では、明確な外在基準と言える指標が存在せず、（EFPA 評価モデルが定義するところの）基準関連妥当性の検証はなされていないため、本稿では構成概念妥当性にのみ焦点を当てる。

11.1.1 では、妥当性の検証に用いられたデザインについて記録する。探索的因子分析、確認的因子分析、項目－合計相関、集団間の構造不変性と差異項目機能の検証、集団間の得点の差、他の尺度やパフォーマンス基準との相関、多特性多方法相関、項目方法理論、実験デザイン、その他の方法のいずれを用いたかが記録される。

11.1.2 では、探索的または確認的因子分析の結果が想定された構造を支持するか否かが評価される。

11.1.3 では、項目得点と検査得点の相関である項目－合計相関を評価する。一般的に相関が高いほど、各項目が指標として有効に機能していることを示すが、高すぎる相関は測定される範囲の狭さを示唆する可能性があるため注意が必要である。

11.1.4 では集団間の因子構造の不変性と差異項目機能が評価される。因子構造の不変性は、複数の集団間で同一の因子構造が保たれることを意味し、項目配置が等しい配置不変、因子負荷量が等しい弱測定不変、切片が等しい強測定不変など、複数の水準が存在する。構造不変性の検証の中で、一部項目の負荷量や切片に集団間で差異が見られた場合、それを差異項目機能と呼ぶ。差異項目機能が見られた場合、それが検査得点に及ぼす影響を推定する必要がある。

11.1.5 では、集団間で想定された平均値の差が見られるかを評価する。例えば、ADHD の診断のある子どもが、診断のない子どもよりも、多動性に関する検査で高い平均値を示すかといった観点で評価が行われる。この種のデータは、疾患や障害の診断やスクリーニングに用いられる検査では臨床的妥当性とも呼ばれ、検査の妥当性を示す最も重要な根拠となりうる。

11.1.6 では、同様の構成概念を測る他の尺度との相関について評価される。これは収束的妥当性と呼ばれ、妥当性検証の不可欠な要素となる。相関の評価は、各検査の測定対象となっている構成概念の類似性によ

って柔軟に行われる必要があるが、典型的な基準として、.55 以上で「適格」、.65 以上で「よい」、.75 以上で「優れている」という基準が設定されている。11.1.7 では、反対に、検査が測定を意図していない構成概念に関する尺度との相関の低さ、つまり弁別的妥当性が評価される。

11.1.8 では、多特性多方法デザインが用いられた場合、その結果が検査の構成概念妥当性を支持しているか否かが評価される。多特性多方法デザインは、複数の特性について、複数の方法（例えば自己評定と保護者評定）によって測定したデータの相関行列を検証する方法である。この方法が用いられる場合、11.1.6 や 11.1.7 の評価は必須ではなくなる。

11.1.9 では、項目反応理論、実験デザインやその他の方法の結果について評価を行う。

11.1.10 では、妥当性の検証に用いられたサンプルのサイズを評価する。古典的テスト理論による検証の場合、100 以上の単一の研究で「適格」、200 以上の単一の研究か 100 以上の複数の研究で「よい」、200 以上の複数の研究で「優れている」と評定される。

11.1.11 では、収束的・弁別的妥当性の評価に基づいて、マーカーとしての検査の質を評価する。

11.1.12 では、妥当性研究が何年前に行われたかが記録される。

11.1.13 では、11.1.1～11.1.12 の評定や記録に基づいて、検査の構成概念妥当性の全体的な適格性を評価する。妥当性研究の結果に加えて、用いられた方法の適切性、検査の対象集団とサンプルの一致、サンプルサイズ、他の尺度の質、研究の時期などについて総合的に評価する。

２. 療育手帳判定に用いられるアセスメント検査の心理測定学的性質

次に、１で紹介した EFPA 検査評価モデルの枠組みに沿って、療育手帳判定に用いられているアセスメント検査の心理測定学的性質について検証する。対象の検査として、過去の実態調査（柴田, 2004; 吉村他, 2019）や本研究班の小林（2021）の分担報告に基づいて、各自治体の療育手帳判定において利用実績のある 9 つの検査を同定した。すなわち、知能検査として、ウェクスラー式検査である Wechsler Preschool and Primary Scale of Intelligence-Third Edition（WPPSI-III）および Wechsler Intelligence Scale for Children－Fourth Edition（WISC-IV）、ビネー式検査である田中ビネー知能検査 V（田中ビネーV）および鈴木ビネー知能検査法（鈴木ビネー）、発達検査として、新版 K 式発達検査（新版 K 式）、遠城寺式乳幼児分析的発達検査法（遠城寺式）、津守式乳幼児精神発達診断法（津守式）、適応行動尺度として、Vineland 適応行動尺度第二版（Vineland-II）、S-M 社会生活能力検査第 3 版（S-M）について検証を行うこととした。

１）基本的特徴

各検査の基本的特徴を表 2 に示す。吉村他（2019）の調査によれば、療育手帳判定に用いられる割合が最も高い検査は田中ビネーであり、K 式がそれに続く。この２つの検査は、対象年齢が乳幼児から成人までと幅広いこと、ウェクスラー式検査に比べて実施の所要時間が短いことなどの実用上の利便性から、療育手帳判定に広く用いられているものと推察される。

しかし、本研究班の内山（2021）の分担報告が指摘するように、国際的診断基準で

は、知的機能や適応行動の水準が一般母集団の平均を 2 標準偏差以上下回ることが知的障害の診断の要件となっている（APA, 2013; WHO, 2018）。これを踏まえれば、知的障害児者を対象とする療育手帳の交付判定には、平均からの偏差を評価できる指数（偏差 IQ など）を用いる必要がある。しかし、上記の 2 検査を始め、国内で利用できる検査の多くは精神年齢と生活年齢の比に基づく比率 IQ（または DQ）を採用しており、偏差指数を利用できる検査はウェクスラー式検査と Vineland-II に限定される。田中ビネーV では、14 歳以上で偏差 IQ を求めることが可能になったが、療育手帳判定の主な対象である乳幼児や児童では比率 IQ しか算出することができない。

このことは、医学的診断基準との乖離という問題に留まらず、判定基準のブレという、より実際的な問題にもつながる。偏差 IQ と比率 IQ では、その定義の違いから、数値の性質にも大きな違いがある。例えば、療育手帳の判定基準を 70 に定めたとしても、偏差 IQ の 70 と比率 IQ の 70 が同じ知的水準を表す保証は存在しない。また、知的能力が年齢とともに直線的に発達するもの対象者の年齢によっても大きく異なることとなる（発達のスピードが速い低年齢ほど、70 の数値が表す意味は重くなる）。

以上より、比率指数に基づく検査を療育手帳の判定に用いることは、医学的診断との乖離を生じさせるのみならず、判定の妥当性をも毀損する重大な原因となりうる。こうした問題について、現場にも広く周知を図る必要があると思われる。

2）基準値

EFPA 評価モデルに基づいて、各検査の基準値について評価した結果を表 3 に示す。表中の数値は、0 が「情報なし」、1 が「不適格」、2 が「適格」、3 が「よい」、4 が「優れている」を表す。新版 K 式については、マニュアルに標準化の手続きに関する記載が見られなかったため、全ての項目を 0 評価とした。EFPA 評価モデルでは、ユーザーが検査の性質を正確に理解し、適切に使用するために、検査に関する情報をユーザーに公開すること自体が検査の質を表す重要な指標であると見なしている。その点で、適切な情報公開がなされていない検査には重大な問題があると言わざるを得ない。

表 2　療育手帳判定の判定に用いられる標準化検査の基本的特徴

	知能検査				発達検査			適応行動尺度	
	WPPSI-III	WISC-IV	田中ビネーV	鈴木ビネー	新版K式	遠城寺式	津守式	Vineland-II	S-M
対象年齢	2:6-7:3	5-16	2-成人	2-18	0-成人	0-4:7	0-7	0-92	1-13
現行版の刊行年	2010	2017	2003	2007	2020	1977	1965・1995	2014	2016
所要時間	40-70分	60-90分	30-60分	35-50分	20-60分	15分	20分	20-60分	20分
指数	偏差	偏差	比率・偏差	比率	比率	（比率）	比率	偏差	比率
療育手帳判定での利用割合[1]	6.50%		51.90%	7.30%	22.60%	9.50%	不明	不明	不明

[1] 吉村他（2019）に基づく。ただし、旧バージョンを含む割合。

でない以上、比率 IQ の 70 が表す意味は、　　9.1.1 については、各地方の人口分布に基

づく割当を行い、代表性の高いサンプルを収集している WPPSI-III、WISC-IV、Vineland-II を 3 評価とした。遠城寺式、津守式、S-M では、人口分布に基づく割当が行われておらず、サンプルが収集された地域の内訳も示されていないため、1 評価とした。その他の検査は、該当する記述が見られなかったため 0 評価とした。

9.1.2 については、療育手帳判定という利用場面に適した偏差指数の基準値が月齢段階ごとに示されている WPPSI-III、WISC-IV、Vineland-II を 4 評価とした。田中ビネーV は、偏差 IQ の基準値が 14 歳以上で示されているが、療育手帳判定の主要な対象となる乳幼児・児童については示されていないため、2 評価とした。その他の検査は療育手帳の判定に適さない比率 IQ の基準値しか示されていないため、1 評価とした。

9.1.3 および 9.1.4 では、十分なサンプルサイズが収集されている WISC-IV、遠城寺式、津守式を 4 評価とし、その他の検査を 3 評価とした。

9.1.5 については、WPPSI-III、WISC-IV、Vineland-II が性別・月齢区分・地域による割当サンプリングを行っていた。残りの検査は、リクルートした対象者全員をサンプルとする簡便サンプリングに基づいていた。

9.1.6 では、上記のような割当サンプリングが行われている WPPSI-III、WISC-IV、Vineland-II では比較的代表性の高いサンプルが得られていると判断し、3 評価とした。その他の検査については、簡便サンプリングを用いており、代表性を示す情報も提示されていないことから 1 評価とした。

9.1.7 では、性別による測定バイアスに関する情報が年齢帯ごとに示されている Vineland-II を 3 評価とした。その他の検査では、集団間の測定バイアスによる検証はなされていなかったため 1 評価とした。

9.1.8 では、標準化が行われた時期によって 10 年以内の 3 検査を 4 評価、10〜15 年前の 1 検査を 3 評価、15〜20 年前の 2 検査を 2 評価、20 年以上前の 2 検査を 1 評価とした。

9.1.9 では、パフォーマンス検査である知能検査について、検査を複数回実施したときの練習効果（得点の上昇）について、年齢

表 3　EFPA 評価モデルに基づく各検査の基準値の評価

		知能検査				発達検査			適応行動尺度	
		WPPSI-III	WISC-IV	田中ビネーV	鈴木ビネー	新版K式	遠城寺式	津守式	Vineland-II	S-M
9.1.1	利用地域に適合したサンプル	3	3	0	0	0	1	1	3	1
9.1.2	利用場面に適した基準値	4	4	2	1	0	1	1	4	1
9.1.3	サンプルサイズ（古典的）					0	4	4		
9.1.4	サンプルサイズ（連続的）	3	4	3	3	0			3	3
9.1.5	サンプル選択の手続き	割当	割当	簡便	簡便	0	簡便	簡便	割当	簡便
9.1.6	集団基準サンプルの代表性	3	3	1	1	0	1	1	3	1
9.1.7	集団、年齢、性別の効果	1	1	1	1	0	1	1	3	2
9.1.8	基準化研究の古さ	4	3	2	2	0	1	1	4	4
9.1.9	練習効果	3	3	2	0	0	n/a	n/a	n/a	n/a
9.3	全体的な適格性	3	3	1	1	0	1	1	3	1

注：表中の数値は、0が「情報なし」、1が「不適格」、2が「適格」、3が「よい」、4が「優れている」を表す。

表 4　EFPA 評価モデルに基づく各検査の信頼性の評価

		知能検査				発達検査			適応行動尺度	
		WPPSI-III	WISC-IV	田中ビネーV	鈴木ビネー	新版K式	遠城寺式	津守式	Vineland-II	S-M
10.2	内的整合性									
10.2.1	サンプルサイズ	4	4	0	3	0	0	0	4	4
10.2.2	報告された係数の種類	折半法	折半法	0	折半法	0	0	0	α	α
10.2.3	係数のサイズ	4	4	0	3	0	0	0	4	4
10.3	検査―再検査安定性									
10.3.1	サンプルサイズ	2	1	2	1	0	0	1	1	1
10.3.2	係数のサイズ	4	4	3	4	0	0	4	4	4
10.3.3	検査―再検査間隔	12-72日	13-86日	6ヶ月	1年	0	0	1-5ヶ月	10-28日	2-41日
10.6	評定者間信頼性									
10.6.1	サンプルサイズ	1	1	0	0	0	0	0	1	1
10.6.2	報告された係数の種類	級内相関	級内相関	0	0	0	0	0	級内相関	相関
10.6.3	係数のサイズ	4	4	0	0	0	0	0	4	3
10.8	全体的な適格性	3	3	2	2	0	0	2	3	3

注：表中の数値は、0が「情報なし」、1が「不適格」、2が「適格」、3が「よい」、4が「優れている」を表す。

帯ごとに詳細な結果が示されているWPPSI-III、WISC-IV を 3 評価、一部の年齢帯の結果のみが示されている田中ビネーV を 2 評価とし、記載がない鈴木ビネーを 0 評価とした。

9.3 では、9.1.1 から 9.1.9 までの評定に基づいて、基準値の全体的な適格性を総合的に評価した。療育手帳判定という利用場面に適した偏差指数を算出することができ、割当サンプリングにより比較的代表性の高いサンプルが収集され、標準化が比較的最近に行われている WPPSI-III、WISC-IV、Vineland-II を 3 評価とした。その他の検査は、利用場面に適さない比率指数による評価を基本としていること、サンプリングの方法が適切でないこと、サンプルの代表性に関する記述も不十分であることから 1 評価とした。

3）信頼性

表 4 に各検査の信頼性の評価の結果を示す。新版 K 式と遠城寺式については信頼性に関する記載がなかったため、全て 0 評価とした。なお、いずれの検査でも検証が行われていない項目については評価しなかった。

10.2 の内的整合性については、WPPSI-III、WISC-IV、Vineland-II、S-M の 4 検査が十分なサンプルサイズのもとで、十分な信頼性係数（.90 以上）を見出していることから、10.2.1、10.2.3 のいずれも 4 評価とした。鈴木ビネーは、サンプルサイズがやや不足し、係数も十分ではなかったため、3 評価とした。その他の検査は検証が行われていないか、情報の記載がなかったため 0 評価とした。

10.3 の検査―再検査安定性の 10.3.1 については、サンプルサイズが 100 名以上であった WPPSI-III と田中ビネーV のみ 2 評価とし、その他の検査は 1 評価とした。10.3.2

表 5　EFPA 評価モデルに基づく各検査の妥当性の評価

		知能検査				発達検査			適応行動尺度	
		WPPSI-III	WISC-IV	田中ビネーV	鈴木ビネー	新版K式	遠城寺式	津守式	Vineland-II	S-M
11.1.2	因子分析	4	4	2	0	0	0	0	2	2
11.1.3	項目一合計相関	0	0	0	0	0	0	0	3	3
11.1.4	構造不変性・差異項目機能	0	2	0	0	0	0	0	2	0
11.1.5	集団間の差	2	2	0	0	0	0	2	4	0
11.1.6	収束的妥当性	3	4	3	2	0	2	2	3	3
11.1.7	弁別的妥当性	0	0	0	0	0	0	0	3	2
11.1.10	サンプルサイズ	2	3	2	1	0	0	1	3	3
11.1.12	妥当性研究の古さ	9年	13年	18年	17年	0	44年	62年	7年	5年
11.1.13	全体的な適格性	2	3	1	1	0	1	1	3	2

注：表中の数値は、0が「情報なし」、1が「不適格」、2が「適格」、3が「よい」、4が「優れている」を表す。

については、.80 以上の係数を示した 6 検査を 4 評価、.80 を下回った田中ビネーV を 3 評価とした。

10.6 の評定者間信頼性については、WPPSI-III、WISC-IV、Vineland-II、S-M の 4 検査で検証が行われていたが、いずれもサンプルサイズが不十分であったため、10.6.1 は 1 評価とした。10.6.3 については、.80 を上回った WPPSI-III、WISC-IV、Vineland-II を 4 評価、.80 をやや下回った S-M を 3 評価とした。

10.8 では、以上の評定を総合し、信頼性の全体的な適格性を評価した。3 つの観点から信頼性の検証が行われ、サンプルサイズに不足があったものの、おおむね十分な結果が得られた WPPSI-III、WISC-IV、Vineland-II、S-M の 4 検査を 3 評価とした。残りの 3 検査は、信頼性の根拠が十分とは言えないものの、最も重要な観点である検査―再検査安定性について検証されている点を考慮し、2 評価とした。

4）妥当性

表 5 に各検査の妥当性の評価の結果を示す。新版 K 式については信頼性に関する記載がなかったため、全て 0 評価とした。なお、いずれの検査でも検証が行われていない項目については評価しなかった。

11.1.2 については、因子分析により当初の仮説が明確に支持されている WPPSI-III、WISC-IV を 4 評価とした。田中ビネーV、Vineland-II については仮説が明確に支持されているとは言えないため、S-M については先行研究に基づく仮説の設定が十分でないため、いずれも 2 評価とした。

11.1.3 については、数値的な情報が示されている Vineland-II および S-M のみを 3 評価とし、その他の検査では検証が行われていない、または数値が示されていないため 0 評価とした。

11.1.4 では、複数の年齢集団での因子分析が行われ、配置不変のみが確認されている WISC-IV および Vineland-II を 2 評価とし、その他の検査では検証が行われていないため 0 評価とした。

11.1.5 では、多様な臨床群におけるスコアプロフィールを示し、検査の臨床的妥当性を検証している Vineland-II のみを 4 評価とした。WPPSI-III、WISC-IV については、原版において詳細な検証がなされてい

表 6　EFPA 評価モデルに基づく各検査の全体的適格性の評価（表 3〜表 5 より抜粋）

		知能検査				発達検査			適応行動尺度	
		WPPSI-III	WISC-IV	田中ビネーV	鈴木ビネー	新版K式	遠城寺式	津守式	Vineland-II	S-M
9.3	基準値ー全体的な適格性	3	3	1	1	0	1	1	3	1
10.8	信頼性ー全体的な適格性	3	3	2	2	0	0	2	3	3
11.1.13	妥当性ー全体的な適格性	2	3	1	1	0	1	1	3	2

注：表中の数値は、0が「情報なし」、1が「不適格」、2が「適格」、3が「よい」、4が「優れている」を表す。

るものの、日本版でのデータが報告されていないため、2 評価とした。津守式は発達遅延児との比較のみが示されているため 2 評価とした。その他の検査では集団間の比較が行われていないため 0 評価とした。

11.1.6 では、他の検査との十分な相関が示されている WISC-IV のみを 4 評価とし、やや相関係数が不十分な WPPSI-III、田中ビネーV、Vineland-II、S-M を 3 評価、さらに根拠が弱い 3 検査を 2 評価とした。

11.1.7 では、複数の尺度との関連から弁別的妥当性を示している Vineland-II を 3 評価、単一の尺度との間で下位尺度間の相関行列を検証しているが、十分な弁別的妥当性の根拠が得られていない S-M を 2 評価とし、検証が行われていない他の検査を 0 評価とした。

11.1.10 では、おおむね十分なサンプルサイズ（200 以上）による検証が行われた WISC-IV、Vineland-II、S-M を 3 評価とし、やや不足した（100〜200）WPPSI-III と田中ビネーV を 2 評価、著しく不足した（100 未満）鈴木ビネー、津守式を 1 評価、記載がなかった遠城寺式を 0 評価とした。

11.1.11 では、収束的・弁別的妥当性の根拠やサンプルサイズに基づき、おおむね十分な結果が得られた WPPSI-III、WISC-IV、Vineland-II を 3 評価、やや根拠が弱かった 2 検査を 2 評価、著しく根拠が不足した 3 検査を 1 評価とした。

11.1.13 では、以上の評定を総合的に踏まえ、妥当性の全体的な適格性を評価した。おおむね十分なサンプルサイズのもとで、多面的な妥当性の根拠が示された WISC-IV および Vineland-II を 3 評価とし、サンプルサイズや結果がやや不十分であった WPPSI-III および S-M を 2 評価、根拠が著しく弱かった残りの 4 検査を 1 評価とした。

5）総合的評価

表 6 に基準値、信頼性、妥当性の各観点に関する全体的な適格性に関する評価を再掲した。

知能検査に関しては、ウェクスラー式の 2 検査（WPPSI-III、WISC-IV）がいずれの観点にも「適格」(2) 以上の評定を得た。これらの検査は、地域の人口分布を考慮した割当サンプリングにより収集された比較的代表性の高いサンプルに基づき、年齢区分ごとの偏差 IQ の基準値を示している。信頼性・妥当性についても多面的に検証が行われ、おおむね十分な結果が得られている。一方で、ビネー式の 2 検査（田中ビネーV、鈴木ビネー）は、比率 IQ による評価を基本としていること、サンプリングの方法が適切でなく、代表性に関する記載も見られないことから、基準値について「不適格」と判断した。また、信頼性については部分的に根拠が示されているものの、妥当性については根拠の提示が不十分であった。

発達検査については、ビネー式検査と同様に、いずれの検査も偏差指数の基準値が示されておらず、サンプルの代表性も保証されていないことから、基準値は「不適格」と判断された。津守式のみ信頼性に関する部分的な根拠が示されていたが、妥当性に関する根拠の提示はいずれも不十分であった。

適応行動尺度については、Vineland-II がいずれの観点にも「よい」（3）の評定を得た。ウェクスラー式検査と同様に、人口分布を考慮した割当サンプリングによる比較的代表性の高いサンプルに基づき、年齢区分ごとの偏差指数の基準値が示されている。信頼性・妥当性についても多面的な評価により、十分な根拠が提示されている。一方、S-M については、ビネー式検査や発達検査と同様に、療育手帳判定に必要な偏差指数を算出することができず、サンプルの代表性も保証されていないため、基準値について「不適格」の判断とした。信頼性については多面的な検証により十分な根拠が提示されているが、妥当性についてはやや根拠が不足している。

以上より、療育手帳判定においては、知能検査としてウェクスラー式の2検査（WPPSI-III、WISC-IV）、適応行動尺度として Vineland-II の利用が推奨される。一方、ビネー式の2検査（田中ビネーV、鈴木ビネー）、3つの発達検査（新版 K 式、遠城寺式、津守式）および S-M については、知的障害の有無と重症度を判断する必要のある療育手帳判定における利用は推奨されない。

3. 知能と適応行動の組み合わせに基づく判定方法に関する検討

1）背景と目的

DSM-5（APA, 2013）や ICD-11（WHO, 2018）などの国際的診断基準では、知能だけでなく適応行動の水準を標準化検査によって測定し、それら2つの指標に基づいて知的障害の診断を行うことが定められている。これは、知能検査によって測定された IQ が、実際の生活場面における困難さを推し量る上で必ずしも十分な指標ではないことを踏まえたものである（APA, 2013）。また、検査の測定値には検査の信頼性・妥当性の範囲内でランダムまたは系統的な測定誤差が含まれるという統計学的事実からも、2つの独立した検査の結果を総合して判断を行うことには合理性があると言える。

しかし、実際の運用において、知能と適応行動という2つの指標をどのように組み合わせて診断を行うかについて、必ずしも一致した指針は示されていない。DSM-5（APA, 2013）では、DSM-IV（APA, 2000）まで定められていた IQ の数値基準の表示がなくなり、「IQ<70」を必須要件としていた従来の方針から、知能と適応行動を総合した、より柔軟な判断を行う方針への転換が図られている。重症度の判定においては、知能ではなく適応行動によって評価を行う方針に変更されたが、これについても具体的な数値基準は示されていない。

ICD-11（WHO, 2018）では、標準化検査によって測定された知能および適応行動が平均をおおむね2標準偏差以上下回ることが診断の要件とされており、重症度の判定においても、平均－2～3標準偏差が軽度、平均－3～4標準偏差が中度、平均－4標準偏差以下が重度・最重度という基準が定められている。しかし、知能と適応行動という2つの指標をどのように総合するかの方法は示されておらず、例えば IQ が 50（中度相当）、適応行動標準得点が 65（軽度相当）であった場合にどのような判定を行うべき

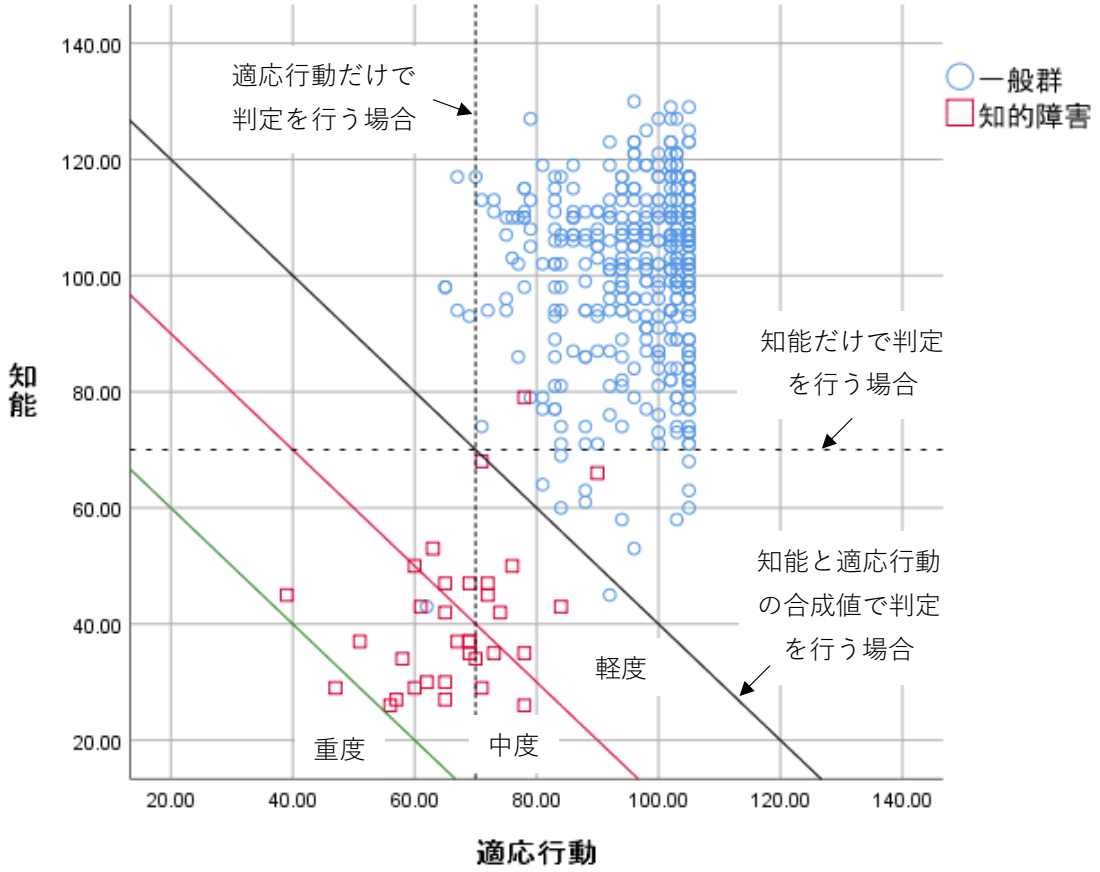

図 1　一般群と知的障害群における知能と適応行動の散布図

かは明確にされていない。

そこで、本稿では一般母集団と知的障害者から得られた知能検査および適応行動尺度のデータを用いて、知能と適応行動をどのように組み合わせて判断を行うことが適切なのか、統計学的観点から検証を行う。ただし、他の事業で収集された成人のデータを用いた予備的検証であり、確定的な結論を得るためのものではなく、今後の研究の方向性を検討するための材料を提供することを目的としたものである。

2）散布図での視覚的検討

図 1 に一般群と知的障害群における知能と適応行動の散布図を示す。いずれの得点も一般母集団の平均が 100、標準偏差が 15 となるように標準化されている。まずはこの散布図において、一般群と知的障害群を最も精度よく判別する方法はどのようなものであるかを考えてみたい。

本研究班の小林（2021）の分担報告によれば、療育手帳判定において、標準化された検査による適応行動のアセスメントを行わず、もっぱら知能検査による IQ によって判定を行っている自治体も多いようである。知能だけで判定を行うということは、図中の縦軸が 70 の位置に水平方向に引かれた破線で判定を行うことを意味する。この場合、知的障害群の大部分が手帳交付の対象となるが、一般群の一部も対象となってしまうことが見て取れる。一方、適応行動だけで判定を行う場合、横軸が 70 の位置に垂直方向に引かれた破線で判定を行うことになる。この場合は、一般群の大部分は交付の対

象とならないが、知的障害群の多くも交付の対象から外れてしまう形になる。

これに対し、知能と適応行動の合成値で判定を行うということは、この散布図に斜め方向の直線を引いて判定を行うことを意味する。２つの指標の重みづけの方法によって直線の傾きは変化するものの、ここでは両者を１：１で合成した値（両者の単純平均値）を用いて判定を行う場合の直線を図示した。この場合、水平方向や垂直方向の直線で区切った場合よりも、一般群と知的障害群を精度よく判別できていることが見て取れる。実際、散布図上での一般群の分布位置と知的障害群の分布位置は縦や横ではなく斜めにずれているため、両者ができる限り重ならないように分けるためには、斜め方向の直線を引くのが合理的であることは明らかである。また、この直線を下方に 30 ずつ平行移動していくことで、軽度、中度、重度の重症度判定の基準となる境界線を得ることも可能である。

３）判別分析

次に、より定量的なエビデンスを得るため、判別分析による検証を行った。判別分析は、複数の量的な独立変数（ここでは知能と適応行動）により、質的な従属変数のカテゴリ（ここでは一般群と知的障害群）を最も精度よく判別する関数を得るための分析である。独立変数として知能のみを含めた場合、適応行動のみを含めた場合、知能と適応行動の両方を含めた場合の結果を表 7 に示した。仮に知能か適応行動のうち一方の変数のみで十分な精度の判別が可能であり、他方の変数が独立の貢献を果たさないとすれば、知能と適応行動の両方を分析に含めたとき、いずれかの判別係数は 0 に近い値を示すはずである。しかし、実際にはいずれも.50 以上の判別係数を示しており、両者がそれぞれ一般群と知的障害群の判別に独立の貢献を果たしていることがわかる。また、判別精度の指標である正準相関係数は、知能のみ、適応行動のみを含めた分析よりも高い値を示している。

これらの結果は、知的障害の診断において、知能や適応行動を単独で用いるよりも、知能と適応行動の合成値を用いることが判定の精度を高めることを示唆している。知能と適応行動は相互に独立した構成概念であることに加え、その測定の方法にも違いがある。つまり、知能検査では課題に対する対象者の反応を記録するのに対し、適応行動尺度では同居者・介護者に普段の対象者の様子について評定を求める。このように、内容と測定方法の両面において異なる指標を組み合わせて用いることで、対象者の実像をより正確に捉え、知的障害者の判別の精度を高めることが可能になっていると考えられる。

判別係数は、個々の変数にどの程度の重みづけをして合成すると判別精度を最大化できるかを表しており、ここでは両者の係数に大きな違いが見られないことから、２）の分析のように、両者の単純平均を用いる形でもおおむね遜色のない判別精度を発揮することができると考えられる。これは実際の運用の利便性を考えたときには、都合の良い性質である（実際、多くの心理尺度では、こうした性質を利用して各項目得点の単純合計を尺度の得点として用いている）。ただし、今回の検討は成人のデータに基づくものであり、実際の運用に適用できる知見を得るためには、乳幼児期から児童・青年期に至るまでの各年齢段階のデータを収集し、判別係数の推定を行うことが必要である。

表 7 一般群と知的障害群の判別分析結果

	判別係数		
	知能のみ	適応行動のみ	知能＆適応行動
知能	1.000		.770
適応行動		1.000	.558
正準相関	.725	.632	.785

D. 結論

本稿では、心理アセスメント検査の体系的な評価の枠組みである EFPA モデルについて概説した上で、療育手帳判定に用いられている 9 つの検査の心理測定学的性質について検証を行った。その結果、知能検査としてはウェクスラー式検査、適応行動尺度としては Vineland-II のみが、療育手帳判定の用途に適合した性能を有していることが示された。他の検査は、知的障害の診断基準にそぐわない比率指数（精神年齢と生活年齢の比）を用いていること、サンプリングの方法が不適切であること、サンプルの代表性に関する情報が提示されていないこと、信頼性・妥当性について十分な根拠が示されていないことなどから、対象者の処遇を左右する high-stakes な判断である療育手帳判定に用いることは適切でないと結論づけられる。

また、本稿では、知能と適応行動という 2 つの独立した指標をどのようにして療育手帳判定という 1 つの判断に利用することが望ましいのか、統計学的観点から予備的検討を行った。成人の一般サンプルおよび知的障害者のデータに基づく検証により、知能、適応行動をそれぞれ単独で用いるよりも、両者の合成値を用いることが知的障害者の判別の精度を高めることが示された。この方法は、手帳交付の判定だけでなく、軽度、中度、重度の重症度の判定においても有効性を発揮すると考えられる。こうした結果に基づけば、現在、多くの自治体で標準化された方法によって適応行動の評価がなされていないことは重大な問題であり、今後、適応行動尺度の普及を図っていくことが必要であると考えられる。

今後の検討課題として、大きく 3 点が挙げられる。第一に、本稿では個々のアセスメント検査のマニュアルに記載された情報に基づいて、各検査の性能について評価を行ったが、これらの検査が実際の運用においてどのように機能し、また、どのような問題を生じさせるのか、実証的な検証が必要である。とりわけ、偏差指数を用いる検査と比率指数を用いる検査の間で、どのような結果の乖離が生じうるのか、体系的な検証が必要である。

第二に、本稿では成人のデータに基づいて、知能と適応行動による知的障害の判別について検証し、両者の合成値を用いることの有用性を確認したが、療育手帳判定の主な対象となる 18 歳未満の対象で同様の結果が得られるかの検証が必要である。

第三に、現在、多くの自治体では所要時間の短さや実施の容易さといった実用上の利便性からビネー式検査や発達検査が広く用いられている現状がある。療育手帳判定にあたる児童相談所の職員が虐待対応など多くの業務を抱える現状で、検査の負担を増やす方向の提言が受け入れられる可能性は高くない。この問題の根本的な解決には、知的障害の有無と重症度の判定という用途に関して十分な心理測定学的性質を持ちつつも、所要時間を大幅に抑えられ、かつ、実施に高度な専門知識が求められないアセスメントツールの開発が必要であると考えられる。我々が並行して進めている社会推進福祉事業「日常生活支援住居施設の対象者選

定のためのシステムに関する調査研究事業」（代表者：辻井正次）では、成人を対象とした簡便かつ高精度の包括的アセスメントツールである ABIT を開発しており、今後、こうしたツールの児童版の開発が求められるであろう。

E. 健康危険情報

該当なし

F. 研究発表

1. 論文発表　なし
2. 学会発表　なし

G. 知的財産権の出願・登録状況

該当なし

H. 引用文献

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders, fifth edition. American Psychiatric Association, Washington D.C.

Bartram, D. (2002a). *EFPA Review Model for the description and evaluation of psychological instruments: Version 3.2. Evaluation Form.* Brussels: EFPA Standing Committee on Tests and Testing (September, 2002).

Bartram, D. (2002b). *EFPA Review Model for the description and evaluation of psychological instruments: Version 3.2. Notes for Reviewers*. Brussels: EFPA Standing Committee on Tests and Testing (September, 2002).

Evers, A., Hagemeister, C., & Hostmaelingen, A. (2013). EFPA review model for the description and evaluation of psychological and educational tests (Tech. Rep. Version 4.2. 6). Brussels: European Federation of Psychology Associations

小林真理子. (2021). 児童相談所および知的障害者更生相談所における療育手帳の判定基準の統一化に向けての課題の整理. 辻井正次. 令和 2 年度厚生労働科学研究費補助金「療育手帳に係る統一的な判定基準の検討ならびに児童相談所等における適切な判定業務を推進させるための研究」報告書.

柴田長生. (2004). 知的障害と発達診断. 発達, 99, 44-51

内山登紀夫. (2021). 現在の知的障害に関する国際的な診断基準と、最近の知的障害概念の検討. 辻井正次. 令和 2 年度厚生労働科学研究費補助金「療育手帳に係る統一的な判定基準の検討ならびに児童相談所等における適切な判定業務を推進させるための研究」報告書.

World Health Organization. (2018). International statistical classification of diseases for mortality and morbidity statistics (11th Revision). https://icd.who.int/browse11/l-m/en

吉村拓馬, 大西紀子, 恵良美津子, 松田裕之, 小橋川晶子, 広瀬宏之, & 大六一志. (2019). 療育手帳判定における知能検査・発達検査に関する調査. *LD 研究*, *28*(1), 144-153.

基準値

集団基準による解釈

9.1　集団基準による基準化

9.1.1　地域での使用に対する適切性、地域の基準か国際的な基準か

n/a　該当せず

0　情報提供なし

1　地域的に関連しない（例：適切でない他国のサンプル）

2　適用地域に適合しないが、注意すれば使用できる地域サンプル

3　適用地域によく関連した地域サンプルか国際的サンプル

4　適用地域のよく定義された母集団から得られた地域サンプルか国際的サンプル

9.1.2　意図した適用に対する適切性

n/a　該当せず

0　情報提供なし

1　意図した適用に適合しない集団基準

2　全体的または部分的には適格な集団基準

3　よい範囲の集団基準

4　年齢および性別に関連し、他の集団差（例：民族集団）に関する情報も示された、優れた範囲の集団基準

9.1.3　サンプルサイズ（古典的な基準化）

n/a　該当せず

0　情報提供なし

1　不適格なサンプルサイズ（Low-stakes な使用：200 未満、High-stakes な決定：200-299）

2　適格なサンプルサイズ（Low-stakes な使用：200-299、High-stakes な決定：300-399）

3　よいサンプルサイズ（Low-stakes な使用：300-999、High-stakes な決定：400-999）

4　優れたサンプルサイズ（Low-stakes な使用：1000 以上、High-stakes な決定：1000 以上）

9.1.4　サンプルサイズ（連続的な基準化）

n/a　該当せず

0　情報提供なし

1　不適格なサンプルサイズ（例：8 未満の下位集団で各 69 以下）

2　適格なサンプルサイズ（例：8 下位集団で各 70-99）

3　よいサンプルサイズ（例：8 下位集団で各 100-149）

4　優れたサンプルサイズ（例：8 下位集団で各 150 以上）

9.1.5　サンプル選択に用いられた手続き

情報提供なし

確率的サンプル―ランダム

確率的サンプル―系統的

確率的サンプル―層化

確率的サンプル―クラスター

確率的サンプル―多相（例：クラスター化した後でランダム抽出）

非確率的サンプル―簡便

非確率的サンプル―割り当て

非確率的サンプル―スノーボール

非確率的サンプル―目的的

その他

9.1.6　集団基準サンプルの代表性

n/a　該当せず

0　情報提供なし

1　意図した適用領域に不適格な代表性または提供された情報から代表性が判断できない

2　適格

3　よい

4　優れている：データがランダム抽出モデルにより収集されている；関連する背景変数（性別、年齢、教育、文化的背景、職業など）に関してサンプルと母集団の構成の徹底した記述が提供されている；これらの変数に関するよい代表性が確立されている

9.1.7　マイノリティ集団間の差、年齢や性別の効果に関する情報の質

n/a　該当せず

0　情報提供なし

1　不適格な情報

2　適格な全体的情報と最小限の分析

3　集団と差のよい記述と分析

4　使用と解釈に関連する重要な問題の優れた分析と考察

9.1.8　基準化研究の古さ

n/a　該当せず

0　情報提供なし
1　不適格（20 年以上前）
2　適格（15-19 年前）
3　よい（10-14 年前）
4　優れている（10 年以内）

9.1.9　練習効果（パフォーマンステストのみ）
n/a　該当せず
練習効果が予想されるのに情報提供なし
全体的な情報の提供
典型的な検査ー再検査間隔を置いた二回目の検査の集団基準の提供

数値基準による解釈

9.2.1　領域基準による基準化

9.2.1.1　臨界得点の決定に専門家の判定が用いられた場合、判定者は適切に選択され、訓練されたか
n/a　該当せず
0　情報提供なし
1　不適格
2　適格
3　よい
4　優れている

9.2.1.2　臨界得点の決定に専門家の判定が用いられた場合、判定者の数は適格か
n/a　該当せず
0　情報提供なし
1　不適格（１名）
2　適格（２名）
3　よい（３名）
4　優れている（４名以上）

9.2.1.3　臨界得点の決定に専門家の判定が用いられた場合、どの基準設定手続きが用いられたか
Nedelsky
Angoff
Ebel
Zieky and Livingston（限定集団）
Berk（対照集団）

Beuk

Hofstee

その他

9.2.1.4　臨界得点の決定に専門家の判定が用いられた場合、評定者間一致を算出するのにどの方法が用いられたか

p0 係数

Kappa 係数

Livingston 係数

Brennan and Kane 係数

級内相関

その他

9.2.1.5　臨界得点の決定に専門家の判定が用いられた場合、評定者間一致係数はどの程度であったか

n/a　該当せず

0　情報提供なし

1　不適格（例：$r < .60$）

2　適格（例：$.60 \leq r < .70$）

3　よい（例：$.70 \leq r < .80$）

4　優れている（例：$r \geq 0.80$）

9.2.1.6　基準化研究の古さ

n/a　該当せず

0　情報提供なし

1　不適格（20 年以上前）

2　適格（15-19 年前）

3　よい（10-14 年前）

4　優れている（10 年以内）

9.2.1.7　練習効果（パフォーマンステストのみ）

練習効果が予想されるのに情報提供なし

全体的な情報の提供

典型的な検査ー再検査間隔を置いた二回目の検査の集団基準の提供

9.2.2　数値基準による基準化

9.2.2.1　臨界得点が実証研究に基づいている場合、その研究の結果と質はどの程度か

n/a　該当せず

0　情報提供なし

1　不適格
2　適格
3　よい
4　優れている

9.2.2.2　基準化研究の古さ

n/a　該当せず
0　情報提供なし
1　不適格（20 年以上前）
2　適格（15-19 年前）
3　よい（10-14 年前）
4　優れている（10 年以内）

9.2.2.3　練習効果（パフォーマンステストのみ）

練習効果が予想されるのに情報提供なし
全体的な情報の提供
典型的な検査ー再検査間隔を置いた二回目の検査の集団基準の提供

9.3　総合的な適格性（9.1-9.2.2.3 に基づいて評定）

n/a　該当せず
0　情報提供なし
1　不適格
2　適格
3　よい
4　優れている

信頼性

10.1　信頼性に関するデータの提供

情報提供なし
単一の信頼性係数のみ提供
単一の測定標準誤差の推定値のみ提供
複数の異なる種類の信頼性係数の提供
複数の異なる種類の測定標準誤差の提供

10.2　内的整合性

10.2.1　サンプルサイズ

n/a　該当せず
0　情報提供なし
1　単一の不適格な研究（例：100 未満のサンプル）
2　単一の適格な研究（例：100-200 のサンプル）

3　単一の大規模な研究（例：200 以上のサンプル）または複数の適格なサイズの研究

4　よい範囲の適格から大規模な複数の研究

10.2.2　報告された係数の種類

n/a　該当せず

α 係数または KR-20

λ -2

最大下限

ω（因子分析）

θ（因子分析）

その他

10.2.3　係数のサイズ

n/a　該当せず

0　情報提供なし

1　不適格（例：$r < .70$）

2　適格（例：$.70 \leq r < .80$）

3　よい（例：$.80 \leq r < .90$）

4　優れている（例：$r \geq 0.90$）

10.2.4　信頼性係数の推定に用いられたサンプル

n/a　該当せず

意図した被検者に適合せず、より有利な係数につながるサンプル（例：人工的な異質性による拡大）

意図した被検者に適合しないが、係数への効果は不明確なサンプル

意図した被検者に適合しないが、より不利な係数につながるサンプル（例：範囲の制限による縮小）

意図した被検者に適合するサンプル

10.3　検査―再検査安定性（時間的安定性）

10.3.1　サンプルサイズ

n/a　該当せず

0　情報提供なし

1　単一の不適格な研究（例：100 未満のサンプル）

2　単一の適格な研究（例：100-200 のサンプル）

3　単一の大規模な研究（例：200 以上のサンプル）または複数の適格なサイズの研究

4　よい範囲の適格から大規模な複数の研究

10.3.2　係数のサイズ

n/a　該当せず

0　情報提供なし

1　不適格（例：r < .60）

2　適格（例：.60 ≤ r < .70）

3　よい（例：.70 ≤ r < .80）

4　優れている（例：r ≥ 0.80）

10.3.3　検査―再検査間隔に関するデータの提供

n/a　該当せず

情報提供なし

時間間隔

10.3.4　信頼性係数の推定に用いられたサンプル

n/a　該当せず

意図した被検者に適合せず、より有利な係数につながるサンプル（例：人工的な異質性による拡大）

意図した被検者に適合しないが、係数への効果は不明確なサンプル

意図した被検者に適合しないが、より不利な係数につながるサンプル（例：範囲の制限による縮小）

意図した被検者に適合するサンプル

10.4　等化信頼性

10.4.1　サンプルサイズ

n/a　該当せず

0　情報提供なし

1　単一の不適格な研究（例：100 未満のサンプル）

2　単一の適格な研究（例：100-200 のサンプル）

3　単一の大規模な研究（例：200 以上のサンプル）または複数の適格なサイズの研究

4　よい範囲の適格から大規模な複数の研究

10.4.2　平行性の仮定は満たされているか

n/a　該当せず

0　情報提供なし

1　不適格

2　適格

3　よい

4　優れている

10.4.2　係数のサイズ

n/a　該当せず

0　情報提供なし

1　不適格（例：r < .70）

2　適格（例：.70 ≤ r < .80）

3　よい（例：.80 ≤ r < .90）

4　優れている（例：r ≥ 0.90）

10.3.4　信頼性係数の推定に用いられたサンプル

n/a　該当せず

意図した被検者に適合せず、より有利な係数につながるサンプル
（例：人工的な異質性による拡大）

意図した被検者に適合しないが、係数への効果は不明確なサンプル

意図した被検者に適合しないが、より不利な係数につながるサンプル
（例：範囲の制限による縮小）

意図した被検者に適合するサンプル

10.5　項目反応理論に基づく方法

10.5.1　サンプルサイズ

n/a　該当せず

0　情報提供なし

1　単一の不適格な研究

2　単一の適格な研究

3　単一の大規模な研究または複数の適格なサイズの研究

4　よい範囲の適格から大規模な複数の研究

10.5.2　報告された係数の種類

n/a　該当せず

推定された潜在特性の信頼性

ρ

情報関数

その他

10.5.3　係数のサイズ

n/a　該当せず

0　情報提供なし

1　不適格（例：r < .70；情報 < 3.33）

2　適格（例：.70 ≤ r < .80；3.33 ≤ 情報 < 5.00）

3　よい（例：.80 ≤ r < .90；5.00 ≤ 情報 < 10.00）

4　優れている（例：r ≥ 0.90；情報 ≥ 10.00）

10.6　評定者間信頼性

10.6.1　サンプルサイズ

n/a　該当せず

0　情報提供なし

1　単一の不適格な研究（例：100 未満のサンプル）

2　単一の適格な研究（例：100-200 のサンプル）

3　単一の大規模な研究（例：200 以上のサンプル）または複数の適格なサイズの研究

4　よい範囲の適格から大規模な複数の研究

10.6.2　報告された係数の種類

n/a　該当せず

一致パーセント

Kappa 係数

級内相関

Iota 係数

その他

10.6.3　係数のサイズ

n/a　該当せず

0　情報提供なし

1　不適格（例：r < .60）

2　適格（例：.60 ≤ r < .70）

3　よい（例：.70 ≤ r < .80）

4　優れている（例：r ≥ 0.80）

10.7　その他の信頼性推定

10.7.1　サンプルサイズ

n/a　該当せず

0　情報提供なし

1　単一の不適格な研究（例：100 未満のサンプル）

2　単一の適格な研究（例：100-200 のサンプル）

3　単一の大規模な研究（例：200 以上のサンプル）または複数の適格なサイズの研究

4　よい範囲の適格から大規模な複数の研究

10.7.2　方法の記述

10.7.3　結果

n/a　該当せず
0　情報提供なし
1　不適格
2　適格
3　よい
4　優れている

10.8　総合的な適格性（10.1-10.7.3 に基づいて評定）
n/a　該当せず
0　情報提供なし
1　不適格
2　適格
3　よい
4　優れている

妥当性

構成概念妥当性

11.1.1　用いられたデザイン
情報提供なし
探索的因子分析
確認的因子分析
（修正済み）項目ー合計相関
構造の不変性と集団間の差異項目機能の検証
集団間の差
他の尺度やパフォーマンス基準との相関
多特性多方法相関
項目反応理論の方法論
（疑似）実験デザイン
その他

11.1.2　因子分析の結果は検査の構造を支持するか
0　情報提供なし
1　不適格
2　適格
3　よい
4　優れている

11.1.3　各項目は検査得点と十分に相関するか
0　情報提供なし

1　不適格

2　適格

3　よい

4　優れている

11.1.4　集団間で因子構造は不変か、差異項目機能はないか

0　情報提供なし

1　不適格

2　適格

3　よい

4　優れている

11.1.5　集団間で予想された平均得点の差があるか

0　情報提供なし

1　不適格

2　適格

3　よい

4　優れている

11.1.6　類似の構成概念の検査との相関の中央値と範囲

0　情報提供なし

1　不適格（$r < 0.55$）

2　適格（$0.55 \leq r < 0.65$）

3　よい（$0.65 \leq r < 0.75$）

4　優れている（$r \geq 0.75$）

11.1.7　検査が測定を意図しない構成概念に関する他の検査との相関は、よい弁別的妥当性を示すか

0　情報提供なし

1　不適格

2　適格

3　よい

4　優れている

11.1.8　多特性多方法デザインが使用された場合、結果は検査の構成概念妥当性を支持するか

0　情報提供なし

1　不適格

2　適格

3　よい

4　優れている

11.1.9　その他（例：項目反応理論、実験デザイン）

0　情報提供なし

1　不適格

2　適格

3　よい

4　優れている

11.1.10　サンプルサイズ

0　情報提供なし

1　単一の不適格な研究（例：100 未満）

2　単一の適格な研究（例：100-200）

3　単一の大規模研究（例：200 以上）または複数の適格な研究

4　よい範囲の複数の適格から大規模な研究

11.1.11　基準やマーカーとしての検査の質

0　情報提供なし

1　不適格な質

2　適格な質

3　よい質

4　収束的・弁別的妥当性に関する広範な指標により示された優れた質

11.1.12　妥当性研究の古さ

年数

11.1.13　構成概念妥当性―総合的な適格性（11.1.1-11.1.12 に基づいて評定）

0　情報提供なし

1　不適格

2　適格

3　よい

4　優れている

基準関連妥当性

11.2.1　基準研究のタイプ

予測的

併存的

事後的

11.2.2　サンプルサイズ

0　情報提供なし

1　単一の不適格な研究（例：100 未満）

2　単一の適格な研究（例：100-200）

3　単一の大規模研究（例：200 以上）または複数の適格な研究

4　よい範囲の複数の適格から大規模な研究

11.2.3　基準測度の質

0　情報提供なし

1　不適格な質

2　適格な質

3　よい質

4　基準構成概念の信頼性と代表性に関する優れた質

11.2.4　検査と基準の関係の強度

0　情報提供なし

1　不適格（$r < 0.20$）

2　適格（$0.20 \leq r < 0.35$）

3　よい（$0.35 \leq r < 0.50$）

4　優れている（$r \geq 0.50$）

11.2.5　妥当性研究の古さ

11.2.13　基準関連妥当性―総合的な適格性（11.2.1-11.2.5 に基づいて評定）

0　情報提供なし

1　不適格

2　適格

3　よい

4　優れている

総合的妥当性

11.3　妥当性―総合的な適格性（11.1.1-11.2.6 に基づいて評定）

0　情報提供なし

1　不適格

2　適格

3　よい

4　優れている