一般化可能性理論を用いた英文和訳問題の採点の信頼性に関する検証

黒川智史 [1]

[1] 東京大学大学院　総合文化研究科　言語情報科学専攻 〒153-0041 東京都目黒区駒場 3-8-1
E-mail: [1] skurokawa0521ut@gmail.com

あらまし　これまで英文和訳の妥当性の検証は行われてきた。その一方、採点方法は統一されておらず、また、採点の信頼性も高いものから低いものまで様々なのが現状である。そこで本稿では、2 人の英語が堪能で、英文和訳に関する独自の評価観を持ち合わせていないと想定される大学院生 2 人に採点のトレーニングを課した後、英文和訳の採点を行ってもらい、その採点の信頼性を検証した。実験には 135 人の高校 3 年生が参加した。採点基準は、複数の大手予備校の大学入試における英文和訳の採点基準を参照しながら作成した。さらに、信頼性の高い採点基準に必要な人数と項目数を求めるために、一般化可能性理論を用いてシミュレーションを行った。その結果、今回のような参加者が 1 時間程度予め採点のトレーニングを行えば､英語が堪能な 1 人の採点者が 2 項目で採点すれば、高い採点の信頼性が得られることが判明した(G = 0.85)。

キーワード　英文和訳，一般化可能性理論，採点の信頼性，採点のトレーニング

Scoring Reliability for Translation Tests Studied with Generalizability Theory

Satoshi KUROKAWA[1]

[1] Department of Language and Information Sciences, Graduate School of Arts and Sciences, the University of Tokyo
3-8-1 Komaba, Meguro-ku, Tokyo, 153-0041 Japan
E-mail: [1] skurokawa0521ut@gmail.com

Abstract　This paper aims to assess translation tests utilizing the Generalizability theory to examine whether they satisfy scoring reliability in Japanese high school term-end exams. In all, 135 Japanese high school students participated in this research. After they completed their translation tests, two well-trained raters assessed each translation task using a translation rubric that was based on the rubric of the most successful cram school and correspondence school in terms of its graduates passing university entrance examinations, the Toshin cram school and the Z-Kai correspondence school. The results indicated that one rater and two items are required to meet scoring reliability in Japanese high school term-end exams (G = 0.85).

Keywords　Translation Tests, Generalizability Theory, Scoring Reliability, Rater Training

1.はじめに

英文和訳は明治の初期から現在に至るまで高等学校の定期テストなどで出題されてきた問題形式である。しかしながら、これまで英文和訳を定期テストで出題することに関して様々な議論があり、それゆえ、英文和訳の妥当性に関する実証的な検証が行われてきた(Buck, 1992; 青木, 2000; Ushiro et al, 2005)。その一方、永原(1936)や馬場(2006)は英文和訳の採点の信頼性に問題があることを指摘したものの、麻生(2012)などを除き、まだあまり検証がなされていないというのが現状である。定期テストにおいて採点の信頼性が確保されていることは、テスト課題の妥当性と同様、非常に重要な要素であり、検証する必要があるといえる。したがって、本研究では英文和訳の採点の信頼性について実証的に検証することにする。

1.1 先行研究

英文和訳の妥当性の検証では、採点方法や信頼性も記されている場合が多い。Buck(1992)では、7 人の短期大学の英語教員に 121 人の大学生が解答した英文和訳を採点してもらい、その採点者間信頼性をピアソンの相関分析を行った結果、最低.087、最高 0.93 という値が算出された。この結果から、Buck は英文和訳の採

黒川智史, “一般化可能性理論を用いた英文和訳問題の採点の信頼性に関する検証”, 言語学習と教育言語学 2018 年度版, pp. 1-9,
日本英語教育学会・日本教育言語学会合同編集委員会編集, 早稲田大学情報教育研究所発行, 2019 年 3 月 31 日.

点者間信頼性は十分あると結論づけた。しかしながら、Buck の実験では、採点者のトレーニングを行っていなかった。青木(2000)では、108 人の参加者の英文和訳を 0 点から 3 点までの範囲で 18 人(大学教員 7 人、高校教員 9 人、中学教員 2 人)が採点した。しかし、英文和訳の採点方法は、具体的な採点基準を設けず、採点者に任せていた。また、Buck 同様、採点者のトレーニングは行っていなかった。Ushiro et al(2005)では、4 人の高校の英語教師が英語の文節ごとに 0 点から 2 点までの部分点を与え、採点者間信頼性をピアソンの相関分析で検証した。その結果、最低 071、最高 0.83 の相関係数が算出された。また、Buck や青木同様、Ushiro et al の実験でも採点者のトレーニングは行われてなかった。

英文和訳の採点の信頼性を中心の課題としたすえた研究としては麻生(2012)がある。麻生は、採点者間信頼性を検証するために、公立高校の英語教師 10 人に採点基準を設けずに 10 点満点と 3 点満点の 2 種類の採点方法で、10 人の高校生が答えた英文和訳を採点してもらった。採点者間信頼性を検証するために、10 人の採点者の採点をそれぞれピアソンの相関分析を用いて分析した。その結果、10 点満点の採点方法では、最低 0.30、最高 0.93 という採点者間同士の相関係数が算出され、3 点満点の場合でも、最低 0.41、最高 0.90 の採点者間同士の相関係数が算出された。このことから、採点者間の相関係数は、10 点満点、3 点満点、どちらのレンジで採点した場合でも、採点者間信頼性は、ばらつきが大きいということが示された。麻生は、点差のレンジにかかわらず、採点にばらつきがてるということを主張している。また、麻生(2012)は、採点者間の信頼性を高めるには、採点のトレーニングが必要であると主張している。

1.2 リサーチクエスチョン

これまでの英文和訳の研究における採点方法は以下の表 1 のようにまとめた。

表1 先行研究の研究結果

	Buck (1992)	青木(2000)	Ushiro et al (2005)	麻生(2012)
採点者の特徴	英語教員	英語教員	英語教員	英語教員
採点者数	7人	18人	4人	10人
トレーニング	なし	なし	なし	なし
点数	5点満点	3点満点	節ごとに0~2点	3点と10点満点
採点基準	教師に委託	教師に委託	教師に委託	教師に委託
分析方法	採点者間信頼性(r)	記述なし	採点者間信頼性(r)	採点者間信頼性(r)
結果	r = 0.87~0.93	記述なし	r = 0.71~0.83	r = 0.30~0.97

このように先行研究では、採点のトレーニングが行われていないこと、そして、評価項目が設けられていないことが示された。他にも、青木(2000)や麻生(2012)のように、多くの研究において採点者は採点のトレーニングが行われていない英語教員のみで構成されていたことから、既に有している英文和訳に対する評価観が存在しており、それが採点に影響を与えている可能性もあることが挙げられる。Weir(2005)によると、スピーキング課題やライティング課題のようなパフォーマンス課題と呼ばれる、採点者の特性や、採点者のトレーニング不足から起因する採点者による採点の厳しさや寛大さは、具体的な採点項目を設けることによって統制を取ることが望ましいという。つまり、本来であれば、日本の英語教育の文脈でも、英文和訳もスピーキング課題やライティング課題と同様に、採点のトレーニングがある方が適切である。

上記の問題点を考慮すると、本研究では、信頼性の高い採点を行うのに必要なトレーニングの内容や、その際に必要な採点者の数、評価項目の数を明らかにすることが考えられる。そうすることで今後の英文和訳の妥当性に関する検証と採点の信頼性に関する研究、また、実際の英語教育の現場にとって、本研究が有意義な検証になると考えられるためである。本来ならば、採点トレーニングの有無、英語教員であるかどうか、評価項目の有無などの要因による英文和訳の採点への影響を検証し、頑強な信頼性の高い英文和訳の採点方法を検証することが研究目的として望ましい。しかしながら、英文和訳の採点の信頼性に関する研究は未だ不足しており、上記の条件をなるべく揃えて採点を行ったとしても、高い採点の信頼性が確保できるかは検証されていない。つまり、たとえ採点のトレーニング、英文和訳に対する評価項目を設けて独自の評価観を持っていないと想定される採点者同士で採点してもらったとしても、あまり高くない信頼性しか得られない可能性もある。その場合、上記の条件以外の要因が採点の信頼性に大きく寄与していると考察されるため、採点に影響するその他の要因を検証しなくてはならないだろう。

そのため、まずは採点トレーニング、英文和訳に対する独自の評価観を持っていないと想定される採点者、採点項目を設けて英文和訳の採点を行った場合、どの程度高い信頼性が確保できるか、また、その際に信頼性の高い採点が可能な採点者の人数、採点項目数を調査する必要があるだろう。そのような検証を行うことで、今後の英文和訳の妥当性の実証研究にあたっては、本研究の採点方法が参考になることが予想される。また、本研究で高い信頼性が得られた場合、今後の英文和訳の採点の信頼性に関する検証において、採点に影響を与える要因(採点者トレーニングの有無、採点者が英語教員なのか、それとも英語教育の現場を経験したことがない大学院生なのか、もしくは異なっているの

か、評価項目の有無)を対象とした発展的な研究に繋がると考察される。したがって、本研究では以下のリサーチクエスチョン(RQ)を立て、検証することにした。

RQ. 英文和訳は、採点のトレーニングを受けた英文和訳に対する独自の評価観を持っていないと想定される採点者が評価項目を設けて採点した場合、何人の採点者と何個の評価項目で採点すれば高い信頼性が確保できるのか

2.実験

2.1 一般化可能性理論

本研究では、RQ を検証するために一般化可能性理論を用いた。一般化可能性理論とは、テスト項目や採点者などの誤差要因によって、得点がどの程度変化するのかを調査するものであり、スピーキング課題などの信頼性の検証において用いられている(小泉, 2018)。他にも一般化可能性理論は、ライティング課題の採点の信頼性に関してもしばしば用いられている(山森, 2002; 山西, 2005; 水本, 2008)。これまで英文和訳において一般化可能理論が用いられた研究は、管見の限り、存在しないが、英文和訳はライティング課題同様、記述式問題であるため、今回の英文和訳の採点の信頼性の検証においても利用できると判断した。

一般化可能性理論は、採点に伴う変動誤差の大きさを分析する「一般化可能性研究（G 研究）」と、G 研究の分散分析の推定値から、一般化可能性係数を算出する「決定研究（D 研究）」で構成されている(小泉, 2018)。一般可能性理論によって、評価項目、採点者、参加者などの変動要因が英文和訳の採点に与える大きさの推定値を求めることができる（山森, 2002; 山西, 2005; 水本, 2008；小泉, 2018）。さらに、山西(2005)は、一般可能性理論によって、設定した目標の信頼係数に到達するのに必要な採点者の人数や、評価項目の数などをシミュレーションすることができると主張している。したがって、本研究の RQ を検証するために、一般化可能性理論を用いて何人の採点者と何個の評価項目で採点を行った場合の信頼係数を算出する。

今回、英文和訳の採点に与える主要因は、「採点者」、「評価項目」、「参加者」とし、系統的誤差を起こす相は 3 相とした。また、主要因を掛け合わせた交互作用は、「参加者」以外の 2 相とした。「採点者×評価項目」、「採点者×参加者」、「参加者×評価項目」とした(山西, 2005)。そして「採点者×評価項目×参加者」には、その交互作用とともに、どの要因をもってしても説明できないランダム誤差が含まれているため、全体で残差を示している(山西, 2005)。一般化可能性理論は、古典的テストの信頼性の概念を発展、拡張させたものであり[1]、そのため D 研究において算出される一般化可能性係数は、古典テスト理論で言うところの信頼係数に相当する(小泉, 2018)。

水本は、山西(2005)を参考にして、何人の採点者が何個の評価項目で採点すれば、ある基準以上の信頼係数が算出できるかを公式にしてまとめた。以下の公式が水本(2008)に記されたものである。

$$G = \frac{\text{参加者の分散成分}}{\text{参加者の分散成分} + \frac{\text{参加者×項目の分散成分}}{\text{項目の数}} + \frac{\text{参加者×採点者の分散成分}}{\text{採点者の数}} + \frac{\text{参加者×項目×採点者の分散成分}}{\text{項目×採点者の数}}}$$

上記の式を用いることで、「(評価)項目の数」と「採点者の数」を変動させ、その採点者数と評価項目数で採点を行った場合の一般化可能性係数をシミュレーションすることができる。また、水本によると、0.8 以上が高い信頼性があるとみなせると主張している。そのため、まずは 0.8 以上の高い信頼係数が得られるまでシミュレーションを行う。0.8 以上に達した後も、より高い信頼性係数である 0.9 以上になるまだシミュレーションを続ける。その理由としては、英文和訳はよりテストの影響度が高い大学や高校の入学試験などでも用いられているため、英文和訳を様々な場面で用いる可能性を考慮したためである。

次に、評価の点数のレンジについて説明する。しばしば、一般化可能性理論はライティング課題の信頼性を検証する場合に用いられる。その場合、項目ごとに、0 点から 5 点まで評価を行うことが多い(山森, 2002；水本, 2008)。ライティング課題は、ある程度以上の長さを有した文章であることから、0 点から 5 点までの項目で評価されることが多いと考えられる。今回の英文和訳は、短い 4 つの文章を採用したため、英文和訳の採点を行った青木(2000)や麻生(2012)同様、各項目 0 点から 3 点までのレンジとした。今回は、後述 2.2 の 4 項目で評価したため、一問あたり最高 12 点、英文和訳は 4 問出題したため、全体で最高 48 点となった。

2.2 英文和訳の採点項目

静(2006)は、大学入学試験に出題されている英文和訳が、高校の英語教育に波及効果を及ぼしていると指摘している。それゆえ、本稿では大学入学試験と関連させて英文和訳の評価項目を検討する。金谷(1995)は、英文和訳の採点項目を決定する際に注意する点について以下のように述べている(pp.217-218)。

[1] 本来は採点方法の改善案を考案するために、G 研究も焦点を当てる必要がある。しかし、管見の限り、英文和訳の採点に一般化可能性理論が用いられていていないため、まずは相対的決定である D 研究に主眼を置いて検証を行った。

テストの妥当性を検証するとき、「タスク」と「採点基準」の設定において、「だれもが納得する」統一基準を設けることが不可能であれば、「一定の判断能力がある人」の中で「より多くの人が納得する」基準を決めるという方法を採るべきである。

(金谷, 1995)

金谷が主張するように、英文和訳では、誰もが納得する判断基準を設定することが困難であるため、本研究では、「一定の判断能力がある人」の中で「より多くの人が納得する」基準を定める方法を設定することにする。

先行研究では、採点者の判断に委託しているため、本研究に応用できる採点基準となりうるものは見当たらなかった(Buck, 1992; 青木, 2000; Ushiro et al, 2005)。それゆえ、教師や生徒が「一定の判断能力がある人」とみなす人が「より多く納得している」と思われる英文和訳の採点基準とは、誰が作成したどのような基準であるかを考察する必要がある。

その考察対象となりうるものとして、大学入試の予備校などの採点基準が考えられる。Allen(2016)は、現在の塾や予備校は、現行の入学試験制度によって生み出された日本の英語教育の「闇の部分(Shadow)」であると述べている。その一方、塾や予備校が、「闇の部分」として機能しているということは、それらが提示している採点基準は、教師や生徒にとって重要な入学試験(High-stakes Test)を解答するために必要な手がかりになっていることを意味しており、それほどまでに塾や予備校は入学試験を徹底的に検証しているといえる。それが教師や生徒が予備校の採点基準を「一定の判断能力がある」とみなしている理由であると考えられる。

また、長きに渡って存在している著名な予備校などは、教師や生徒がその採点基準を「より多くの人が納得している」ため存続し続けていられると考えられる。したがって、本稿では、予備校の採点基準を参照することにする。採点基準を参照するにあたって、1 つの予備校ではなく、複数の伝統的な予備校を参照することにする。

例えば、大手予備校の東進が作成した『第 1 回 6 月東北大本番レベル模試(2013 年 6 月実施)[2]』によると、以下のような英文和訳の採点基準が設けられている。

1.基本的に部分的誤りと語句の無視は各-2 点
2.各採点区分内の配点を超えて減点は行わない
3.十分に日本語化しているカタカナ語は不問だが、その場合を除くカタカナ語は-2 点
4.日本語の誤字等は，別の意味にとられてしまうケースは-1 点

(東進, 2013)

上記の例から読み取れるように、基本的に英語に関する能力を測定しているが、日本語の誤字や別の意味にとられてしまうような表現が減点対象であることから、日本語の能力も問われている、と東進は分析している。さらに永原(1936)同様、部分的な語句のミスは減点対象となみなしている。

もう 1 つ大手通信教育の予備校の英文和訳の採点基準を取り上げる。実は、英語の通信教育には長い歴史がある。江利川(2011)によると、日本の英語教育においては、通信教育の始まりは 1924 年の井上通信英語教育まで遡ることができるという。現在、大手通信教育といえば、Z 会の名前をあげる人も多いだろう。Z 会が 2016 年度の京都大学の英文和訳の勉強方法を示したものを以下に引用する[3]。

英文和訳では、まず英文の構造を文法・構文に基づいて丁寧に分析する練習を積むこと。そして自然な日本語になるよう工夫を重ねる。語彙については、単語集などでこまめに覚えていくと同時に、文脈から未知語の意味を推測する練習が必要である。背景知識や日本語の語彙力も助けになるだろう。

(Z 会, 2016)

このように、英文和訳では、英文の構造について丁寧に分析すること、そして自然な日本語であることが重要であると Z 会は分析していると思われる。

今回取り上げた英文和訳の採点基準でわかったこととは、英語の文構造や、単語だけでなく、与えられた分脈から逸脱した訳は減点対象になること、日本語関連のミス（誤字、脱字、定着していないカタカナ語の使用）なども減点されるということである。当然ながら、大手予備校が作った採点基準がどこまで本来の採点基準と近いものなのかは調べようがない。しかしながら、毎年多くの受験生がこのような採点基準を参考にしてテストに望んでいる。そのため、実際の大学入試に向けて同様の採点基準を設けることは現実的であると考えられる。そして、少なくとも、青木(2000)や麻生(2012)の実証実験のように、細かい採点基準を設けず、教師に任せて英文和訳を評価したものよりも、

[2] 第 1 回 6 月 東北大本番レベル模試(2013 年 6 月実施)より引用した。URL(https://www.toshin.com/hs/event/tohokudai_honban/pdf/1306/eigo.pdf#search=%27英文和訳+採点基準%27 2018 年 11 月 2 日 閲覧)。

[3] 「京大英語 再現答案 & 得点情報分析 『目標得点を達成するためのポイント」より引用した URL(https://www.zkai.co.jp/high/saigen/contents/base.aspx?cd=16ke_b 2018 年 12 月 21 日 閲覧)。

高校の英語教育の現状に即した正確な評価ができると考えられる。

今回の評価の項目は、「英文の文構造を理解しているか」「英文の文章を理解しているか」「日本語の自然さ」「日本語の正しさ」の 4 項目である。「英文の文構造を理解しているか」という項目は、永原(1936)や、大手予備校である東進(2013)が言及した項目であるため取り入れた。「英文の内容を理解しているか」については、馬場(2006)が述べているように、英文和訳はさほど難しくない構文と語彙であった場合、原文の理解度を測定できるとされているため項目として採用した。「日本語の自然さ」と「日本語の正しさ」については、英文和訳は日本語としての能力も必要であると考えられるため、取り入れた。

なお、実際の課題で用いた採点表や、採点事例の一部を Appendix.1 に提示しているので参照されたい。

2.3 実験手順

今回の実験では、参加者に時間内に 4 問の英文和訳を解いてもらった。参加者の答案を全て Excel に転記し、参加者が特定できないようにするために通し番号を振った。そして、採点のトレーニングを受けた 2 人の採点者にその Excel ファイルを印刷したものをそれぞれに渡し、採点してもらった。その後に、採点した結果を一般化可能性理論の G 研究と D 研究を行った。

なお、分析に使用したソフトウェアは、SPSS Advanced Model の VARCOMP であった。

2.4 参加者

参加者は、首都圏にある私立大学付属の高校 3 年生、計 145 人であった。参加者には、2018 年 5 月下旬から 6 月中旬の授業時に英文和訳の課題と英検などの英語の民間試験の取得状況を尋ねる質問紙に答えてもらった。なお、調査対象校は男女別学の男子校であったため、全ての参加者の性別は男性であった。参加者の 9 割近くが大学受験をせずに、付属の大学に進学する。また、付属の大学に進学するためには、英検 2 級かそれと同等と学校が認めた試験に合格しなくてはならない。そのため、いわゆる進学校のように、大学受験に特化した英語教育ではなく、英検 2 級の取得を推奨し、コミュニケーション能力を伸ばすことがこの高校の教育理念の 1 つとして掲げられている。また、帰国子女が複数人在籍しているその一方で、中学まで日本の公立学校に在学していた生徒もいることから、様々な英語教育の経験を有している生徒がいるといえる。本調査では、無記名であったものや、欠損値があった 10 人を除いた 135 人を分析対象とした。なお、収集した全てのデータは、Excel に打ち込む際に、参加者に番号を振り個人が特定できないようにした。Excel にデータを打ち込んだ後も調査で参加者が直接記入した質問紙は自宅で厳重に保管している。

2.5 採点者

英文和訳の採点は、英語が堪能である 2 人の大学院生が行った。今回、大学院生に依頼した理由は、参加者数が 135 人、それぞれ 4 問解いてもらったため、中学や高校の教員に採点してもらうには、英文和訳の数が膨大すぎたためである。そして多忙な中学、高校の教員に採点のトレーニング講習を受けてもらうことが困難であると判断したためである。そこで本研究では、これまで英文和訳の採点をしたことのない大学院生に採点してもらうことにした。採点期間は、2018 年 8 月 4 日から 14 日の間に採点してもらった。

この 2 人は同じ都内の大学院に所属している。両者とも中学・高校の外国語科の教員免許は持っておらず、研究分野も外国語教育関連のものではない。大学院の研究分野はそれぞれ認知言語学(以下: 採点者 A)と、コミュニケーションストラテジー研究(以下: 採点者 B)であった。採点者 A は日本語母語話者、採点者 B は中国語と日本語の均衡バイリンガルであった。

本研究で英語教育分野の研究とは関係がなく、教員免許状を取得していない大学院生の採点者を選んだ理由について説明する。一般化可能性理論では、採点者を母集団からランダムサンプリングで選ばれることになっている。それゆえ、本来であれば、実際に英文和訳を採点していると想定される英語教員が母集団として推定されるため、英語教員が採点することが望ましい。事実、これまでの先行研究においても、英語教員が英文和訳の採点を行っている(青木, 2000; Ushiro et al, 2005)。今回、英語が堪能であるが、英語の教員免許を取得していない大学院生に採点を依頼した理由は、1.2 で述べたように、英語教員や教員志望の学生が評価を行う場合、今までの学習・指導の経験に基づく独自の評価観点があり、それが採点に歪みを生じさせる可能性があるためである。また、英語教員に比べて、英語教育に対する強いこだわりがないと想定されるため、採点のトレーニングによって評価基準に合わせることも容易である可能性があるというのも理由の 1 つである。ゆえに、今回の採点者は、本研究の「英文和訳に対する独自の評価観を持っていないと想定される採点者」として適任であると判断した。

次に、採点者の人数について説明する。先行研究では、英文和訳が 4 人(Ushiro et al, 2005)から 18 人(青木, 2000)で採点されているのに対し、本研究の採点者の人数である 2 人はかなり少ない。多くの採点者がいた方が、より正確な信頼性係数を算出することができるため、採点者の人数をある程度多くすることが望ましい。しかしながら、本研究では、英語が堪能であるものの、英文和訳に対する独自の評価観を持っていないと想定

される採点者が、評価することになっていることから、採点者の人数を確保するのが困難であった。そのため、最も少ない人数で一般化可能性理論を行っているライティング課題の研究を調査した。その結果、ライティング課題の採点の信頼性を一般化可能性理論で検証した水本(2008)では、採点者は 2 人のみで行っていた。そのため、本研究では、水本における人数を一般化可能性理論で検証する際の最小限の人数とみなし、検証することにした。

2.6 採点のトレーニング

両採点者には、2018 年 8 月 5 日の午前 11 時から 1 時間、筆者が主催した採点のトレーニング講座に参加してもらった。用いた採点項目は、2.2 で言及したものと同様である。予備調査として、大学院生が解答した本調査で用いたものと同じ 4 つの英文和訳を設問ごとに 3 問ずつ、計 12 問英文和訳を採点してもらった。トレーニングの際は、練習問題の 12 問の採点した値が一致するまで筆者を交えた 3 人で話し合った。そして 12 問の練習問題の採点が一致したところでトレーニングは終了した。なお、採点者にはそれぞれ、すべての参加者の英文和訳を渡し、2018 年 8 月 5 日から 8 月 13 日の間に採点してもらった。

2.7 英文和訳

英文和訳は、英検のテキストから引用した。具体的には、公益財団法人日本英語検定協会の 2017 年度の第 2 回英語検定一次試験（準 2 級）の長文読解“A Children’s Toy”の文章からに下線を引いた 4 つの文を実験に用いた[4]。その理由は、英検のウェブサイト(2018)にある「各級の目安[5]」によると、英検準二級は「大学入試レベル」と記載されており、さらに、多くの参加者が英検 2 級の習得を目指している。しかしながら、質問紙で資格試験の取得状況を確認したところ、調査時点において英検 2 級以上の資格を取得している参加者は全体の 37%であり、英検準 2 級の一次試験の合格者は 35%であった。また、残りの 28%は英検準 2 級の一次試験を突破していなかった。それゆえ、参加者の上位 3 割程度にとって簡単な英文レベルであり、中位 3 割によっては、適切な英文レベル、下位の 3 割弱にとっては難しい英文レベルであると判断できる。したがって、この課題の英文レベルは参加者にとって適切なレベルであると判断した。

ここで、学校で普段用いている教科書などから既習の長文を引用すべきではないかという批判があるかもしれない。実際に英文和訳が期末テストで使用される際には、これまでの学習事項が定着しているかを確認するために出題していることが多いためである。しかしながら、本稿では、普段使用している教科書のものを使用すると、学習者の読解力を測定しているよりも、教科書を暗記していることを測定してしまう恐れがあるため、教科書からの引用は避けた。

[4] 下記の URL にある公益財団法人日本英語検定協会より引用した(http://www.eiken.or.jp/eiken/exam/grade_p2/solutions.html 2018 年 12 月 21 日 2018 閲覧)。

[5] 下記の「各級の目安」を参照されたい（http://www.eiken.or.jp/eiken/exam/about/ 2018 年 12 月 21 日閲覧）。

3.結果

3.1 G 研究

まず、英文和訳の採点における変動要因である採点者、評価項目、参加者の分散分析の推定値と、各項目の推定値を百分率にしたものを以下の表 2 にまとめた。

表 2 一般化可能性係数研究(G 研究)の検証結果

	分散成分	
変動要因	推定値	□□割合（%）
参加者	6.77	66.4%
採点者	-0.04*	0.00%*
項目	1.109	10.9%
参加者×採点者	0.13	1.2%
採点者×項目	1.305	12.8%
採点者×項目	0.108	1.1%
参加者×採点者×項目	0.773	7.6%
合計	10.2	100%

Note；採点者の項目は分散がマイナスだが、Brennan(1992)の方法に従って、割合は 0%としている。

この表でまとめられているのは、一般化可能性理論における G 研究に該当する。表 1 を見てみると、最も大きい変動要因は「参加者」であった(66.4%)。これは参加者によって評価が異なるということを意味し、参加者の能力の識別ができているといえる。最も変動要因として小さかったのは、「採点者」であった。Brennan(1992)によると、採点者の分散成分がマイナスであり、且つ、-0.05 以内である場合は、その分散の真の値が 0 に近いため発生しているという。英文和訳の採点に与えた要因は限りなく小さくなった要因であることが考察される。したがって Brennan の主張に従い、負の値を 0 に修正した。しかしながら、推定結果の偏りを防ぐため、分散分析においては、マイナスの値をそのまま使用した。「採点者」の変動要因が小さいことは、採点者同士の採点のズレも小さい可能性がある。そこで、先行研究と同様、採点者間信頼性があるのかを採点者 A と採点者 B の 4 項目の点数を足し合わせた英文和訳の得点をピアソンの相関分析を用いて検証した。その結果、r = 0.95 という相関係数が算出された。このことから、採点者 A と B の採点者間信頼性は非常に高いことが示された。

その一方で、「評価項目」の分散要因の大きさが

10.9%あるということは、項目によって測定しているものが大きく異なるということを表している。

3.2 D 研究

一般化可能性理論の次の段階である D 研究に移行する。上記の G 研究から得られた分散推定を用いて、一般化可能係数(G)を求めた。また、採点者数と評価項目数を変動させると、どの程度の信頼係数が算出されるかをシミュレーションする。D 研究における信頼性は、古典的テストでいうところのクロンバックのα係数と同様に扱っている研究は多い(山森, 2002；山西, 2005；水本, 2008)。本調査は決定係数が 0.8 以上の場合に信頼性の高い採点であるとみなすことにする。本調査は採点者 A と採点者 B にそれぞれ同じ英文和訳を 4 項目で採点してもらったため、まず採点者数が 2 人の場合のシミュレーションを表 3 にまとめた。表 3 によると、2 人の採点者が 2 項目で採点を行った場合は 0.91 という決定係数が算出された。また、実際に今回の採点方法であった 2 人の採点者が 4 項目で採点を行った場合 0.93 と高い決定係数が算出されていた。

表 3 採点者数が 2 人の場合のシミュレーション

採点者数	項目数	一般化可能係数
2	1	0.79
2	2	0.88
2	3	0.91
2	4	0.93

また、表 4 は採点者が 1 人の場合のシミュレーションを行ったものである。

表 4 採点者が 1 人の場合のシミュレーション

採点者数	項目数	一般化可能係数
1	1	0.75
1	2	0.85
1	3	0.89
1	4	0.91

表 4 から読み取れるように、採点者が 2 人の時と同様、採点者が 1 人の時においても 2 項目以上で採点すれば決定係数が 0.85 であり、1 人の採点者が 4 項目で採点した場合は 0.91 という高い信頼性があるということが示された。次に、採点者が 1 人の場合と採点者が 2 人の場合を視覚的にわかりやすく比較するために、以下のような図 1 を作成した。

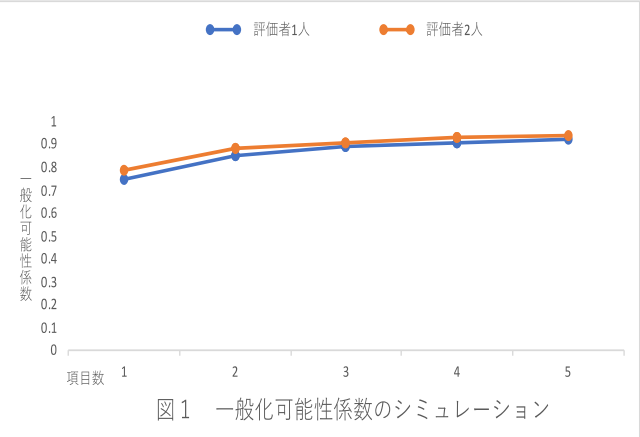

図 1　一般化可能性係数のシミュレーション

図 1 で示されているように、採点者が 1 人でも 2 人でも、評価項目数が 4 項目程度になるとあまり採点の信頼性に大きな差がない(採点者 1 人の場合、決定係数 = 0.91：採点者 2 人の場合、決定係数 = 0.93) ことが明らかになった。

4.考察

まず、一般化可能性理論の G 研究の結果によると、全ての変動要因の中で最も大きかった分散成分は、「参加者(推定値 = 6.77、推定値全体における割合 = 66.4 %)」であったことから、今回実験に用いた英文和訳の課題は、参加者の能力を十分に識別できていたといえる(水本, 2008)。また「採点者」はほとんど英文和訳の点数の変動要因とならなかった(分散成分の推定値 = -0.04 推定値全体における割合 = 0.00 %)。これは、各評価項目が 0 点から 3 点までの比較的小さいレンジであったことを考慮しなければならないが、採点者によってのばらつきはほとんどなかった、ということを意味する。また、採点者 2 人の採点者間信頼性は、r = .95 であったことも、この結果を裏付けるものといえる。同様の方法で信頼性係数を算出した先行研究において、最も強い採点者間信頼性の相関係数は、麻生(2012)の r = 0.97 であったとから、本研究は他の英文和訳を採点した研究と比較しても、かなり強い採点者信頼性が算出されたといえる。つまり、今回のような採点者の場合、厳格な評価項目表を与え、1 時間程度トレーニングすれば採点がかなり一致するということが示唆された。

「採点者×項目」の分散成分の推定値が小さかったことから(推定値 = 0.10、推定値全体における割合 = 1.1 %)、採点者によって項目の難易度が変わることはほとんど起こらないことが示された。

しかしながら、「参加者×項目」の推定値も高かった(推定値 = 1.30、推定値全体における割合 = 12.8%)。これは、少なくとも 1 つの項目が、ある参加者には有利に働いたり、ある参加者に不利に働いたりしていることを示している(Bachman, 1990)。例えば、項目は、主に英語の能力に関連するものと、日本語の能力に関連するものがあるため、「英語の項目で高い点数が取れているにもかかわらず、日本語の項目であまり高い点数が取れなかった参加者」もしくは、「日本語の項目では、高い点数が取れたのにもかかわらず、英語の項目では高い点数を取れなかった参加者」がいたことが想定される。今回は予備校の採点表を参考にして観点別の評価を作成したが、今後の研究では日本語に関する評価を入れない場合どのように採点の信頼性に影響が出るかに関して検討の余地が残されており、今後検証していく必要があるだろう。また、そもそも日本語の評価項目を、英語の課題である英文和訳に入れること自体に関して慎重に議論を重ねる必要もあるだろう。

次に、D 研究の結果を考察する。水本(2008)がライティング課題で信頼性が高いと主張できる目安として 0 信頼性係数が 0.8 を超えることを挙げている。英文和訳の採点の信頼性を考察した本研究で水本の目安を取り入れ、D 研究の決定係数が 0.8 を超えるには、1 人の採点者が 2 項目で行うことが条件であった(決定係数 G = 0.85)。さらに、より高い信頼性を高めるには、採点者は 1 人の場合、4 項目以上が望ましいということが示された(決定係数 G = 0.91)。

上記の考察から、RQ「英文和訳は、採点のトレーニングを受けた英文和訳に対する独自の評価観を持っていないと想定される採点者が評価項目を設けて採点した場合、何人の採点者で何個の評価項目で採点すれば高い信頼性が確保できるのか」に答える。その条件として、(1)英文和訳に対する評価観がないと想定される大学院生を採点者として選び、(2)採点項目を理解してもらい、採点者間の評価方法と評価項目ごとのコンセンサスを取るために1時間で12問程度の採点者に対する採点トレーニングを行うことである。その条件がクリアされた場合、1 人の採点者が、には 2 つの評価項目で採点を行うと 0.8 以上の信頼性係数が得られる。また、0.9 以上の信頼性係数を得るためには 1 人の採点者が 4 項目以上で採点することで得られることが示された。従来の英文和訳の妥当性の研究では、採点の信頼性は、あまり重視されてこなかったため、信頼性係数は研究によって高いものから低いものまで様々である、また、採点方法自体もその研究独自の手法で行われていることも多く、統一された方法で行われているわけではなかった(Buck, 1992; 青木, 2000; Ushiro et al, 2005)。英文和訳の検証において用いることができる採点方法 1 つを示すことが一連の英文和訳の研究に貢献できた部分であると考察される。本研究の限界や、今後検証すべき課題に関しては次の章で述べることにする。

5.今後の展望

本研究では、一般化可能性理論を用いて、採点者のトレーニングを受け、英文和訳に対する独自の評価観を持っていないと想定される採点者が、評価項目を設けて採点した場合、信頼性の高い採点に必要な採点者数と評価項目数を検証した。それを検証するために、英文和訳の採点項目を設けた状態で、1 時間程度の採点トレーニング受けた英文和訳に対する独自の評価観を持っていないと想定される大学院生 2 人に採点してもらい、一般化可能性理論を用いて分析した。その結果、1 人の採点者が 2 つの項目で採点を行った場合、0.8 以上の信頼性係数が算出された。また、1 人の採点者が 4 つの評価項目で評価を行った場合、0.9 以上の高い採点の信頼性係数が算出された。このことから、今回と同様の条件で採点を行った場合、採点の信頼性を確保できることが示された。

その一方で、本研究には様々な限界がある。例えば、今回の採点者は、英文和訳に対する独自の評価観を持っていないと想定される大学院生であったため、今回のトレーニングを、独自の評価観を持っていると想定される英語教師に実施した場合、今回と同様の結果が得られるかどうかは不明なことなどである。最後に今後の英文和訳の採点の信頼性の検証で行う必要があるものを以下に挙げることにする。

(1)採点者のトレーニングの前後でどの程度採点間信頼性や信頼性係数が向上するのか

(2)採点者が英語教師の場合と大学生・大学院生の場合で採点の信頼性に差はあるのか

(3)本実験に参加してもらった採点者 A と B に同様の実験を行い、今回の実験結果と比較し、どの程度採点者内信頼性があるのか

(4)一般化可能性理論の G 研究おいて、日本語の項目を取り除いた場合、「参加者×項目」の分散成分はどのように変化するのか

上記の(1)から(4)を検証することで、頑強な英文和訳の採点方法の考案に繋がるだろう。したがって、今後(1)から(4)の検証を続けていくことが期待される。

文　献

[1] G. Buck, Translation as a language testing procedure: does it work?, *Language Testing*, vol.9, no.2, pp.123-148, December 1992.

[2] 青木優子, “英文和訳テストの妥当性調査, ” 関東甲信越英語教育学会研究紀要, vol.14, pp.35-42, 東京, July 2000.

[3] Y. Ushiro, Y. Hijikata, M. Shimizu, Y. In’nami, K. Kasahara, A. Shimoda H. Mizoshita, & R. Sato, Reliability and Validity of Translation Test as a Measurement of Reading Comprehension, *Annual Review of English Language Education in Japan*, vol.16, pp.71-80, Gunma, May 2005.

[4] 永原敏夫, 試験と学修, 研究社, 東京, December 1936.

[5] 馬場哲生, “英文和訳テストの功罪,” 英語青年, vol.152, no.7, pp.408-410, October 2006.

[6] 麻生雄治, “英文読解力評価のための英文和訳テストの信頼性と妥当性,” EIKEN BULLETIN, vol.24, pp.189-197, November 2012.

[7] C. J. Weir, *Language Testing Validation: An Evidence-Based Approach*, Palgrave Macmillan, Hampshire, November 2005.

[8] 小泉利恵, “一般化可能性理論,” 平井明代 (編著), 教育・心理・言語系研究のためのデータ分析―研究の幅を広げる統計手法, 東京図書, 東京, December, 2018.

[9] 山森光陽, “一般化可能性理論を用いた観点別評価の方法論の検討,” Step Bulletin, vol.14, pp.62-70, November 2002.

[10] 山西博之, “高校生の自由英作文はどのように評価されているのか－分析的評価尺度 と総合的評価尺度の比較を通しての検討－,” JALT Journal, vol.26, pp.189-205, November 2004.

[11] 水本篤, “自由英作文における評定者評価の種類と信頼性,” 統計数理研究所共同研究リポート 215「学習者コーパスの解析に基づく客観的作文評価指標の検討」, vol.215, pp.43-49, March 2008.

[12] 静哲人, “これでいいのか、大学入学英語問題−英語教育およびテスト理論の立場から,” 英語青年, vol.152, No.1,pp.2-5, April, 2006.

[13] 金谷憲, 英語リーディング論, 河源社, 東京, September 1995.

[14] D. Allen, “Japanese cram schools and entrance exam washback,” *The Asian Journal of Applied Linguistics*, vol.3 No.1, pp.54-67, March 2016.

[15] 江利川春雄, 受験英語と日本人:入試問題と参考書からみる英語学習史, 研究社, 東京, March 2011.

[16] L. F. Bachman, *Fundamental considerations in language testing*, Oxford University Press, Oxford, June 1990.

[17] R. L, Brennan, *Elements of generalizability theory*, ACT Publications, Iowa City, December 1992.

Appendix.1

ここでは紙幅の都合上、英文和訳の第 1 問で用いた採点基準を記すことにする。

問題(1) Even with good ideas and hard work, many products and businesses fail.

訳例:たとえ良い発想やハードワークがあっても、多くの商品とビジネスは失敗する

(1)の採点項目表

内容	英文の文構造を理解しているか	英文の文章を理解しているか	日本語の自然さ	日本語の正しさ
◎右の例を正解の参考にしてください	・Even with～ ～ても/ ～にも拘らず/ ～しても/～さえ（も）/ ～でさえ ・～fail 失敗する（自動詞）/うまくいかない	・hard work 大変な仕事/一生懸命努力/懸命な働き/ハードワーク/熱心な仕事/きつい仕事/とても働く ・Products 製品/商品/プロダクト(ツ)/生産物 ・Businesses ビジネス/仕事/会社/事業	文脈を考慮すると、とても自然である ・例；fail（失敗する企業も少なくない）	・漢字がわからないために、平仮名で書いたと思われるものがない ・助詞が重なったりしていない （例；多くの製品がが） ・読点や句点については対象外
△ 1個につき1点ランクダウン	・Even with～ ～で/～すると/～より ・～fail 失敗させる/失敗/失敗をともなう/を失敗する（他動詞）	・hard work 忙しい仕事/沢山の仕事/努力/勤勉/一生懸命研究しても ・products 事業/計画/生産 ・～fail 影響する/厳しいものだ ・ビジネスや企画ができるよい働き（good は idea のみ修飾）	良いアイデアや働きでさえ～ （並列の部分が変） ・例え良い考えがあったり勤勉に働いても、 （勤勉には働かない） 良いアイデアでも、勤勉でも/良い考えやいそがしい仕事で、(二つの関係がよくわからない)	はげしく働く 良いアイデアと、けん命な働きにもかかわらず、 かん単/はっそう ・がんばってはたらいたとしても （がんばって－1/はたいたとしても－1）
×(1個につき2ランクダウン) ＊各項目0点以下にはしません	・Even with～ ～が、～は(主格) ～しなければときで～ （体言止め）	・hard work 熱心な勉強 しっかりと仕事ができる	どんなにグッドアイデアやハードな努力があっても、 （まだ日本語にできる部分がある）	

実際の採点例

(1) Even with good ideas and hard work, many products and businesses fail.	文構造	意味	自然さ	正しさ	合計点
優れた考えや大変な労力があったとしても多くの製品や経営は失敗してしまう	A:3 B:3	A:3 B:3	A:3 B:3	A:3 B:3	A:12 B:12
良い発想やねっしんな仕事をしたが、多くの製品やビジネスは失敗する	A:2 B:1	A:3 B:3	A:2 B:2	A:2 B:2	A:9 B:8
忙しい時に良いアイデアがある	A:0 B:0	A:0 B:0	A:0 B:0	A:1 B:0	A:1 B:0

Note: A、B はそれぞれ採点者 A、B を指し、数字はそれぞれの採点を表している。