社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

GitHubにおけるバグ報告等の動画及び画像の活用実態に関する調査

蔵元　宏樹† 　石本　優太† 　新堂　　風†
近藤　将成† 　柏　祐太郎† 　亀井　靖高†
鵜林　尚靖†

† 九州大学
E-mail: †{kuramoto,ishimoto,shindo}@posl.ait.kyushu-u.ac.jp,
††{kondo,kashiwa,kamei,ubayashi}@ait.kyushu-u.ac.jp

あらまし　バグの症状の迅速で的確な把握は，デバッグ作業の効率に大きく影響する．バグの再現動画やバグの状況を表す画像は，デバッグ作業に有用であると期待できる．本研究の目的は，ソフトウェア開発現場におけるバグ報告動画及び画像の活用実態を明らかにすることである．GitHubで公開されている約4,000件のリポジトリを対象に，課題報告（Issueレポート）において動画及び画像の有無による影響を調査した．その結果として，動画及び画像が含まれる課題報告は，そうでないものに比べて，課題報告が解決するまでの時間が平均で，2.9%~17.6%増加し，またコメント数が平均で76.4%~143%増加した．一方で，最初のコメントがつくまでの時間が平均で20.1%~25.4%減少した．これらの結果の詳細について報告する．

キーワード　GitHub，動画，画像，バグ報告，Issue

The investigation of the usage of videos and images for bug reports in GitHub

Hiroki KURAMTO†, Yuta ISHIMOTO†, Kaze SHINDO†,
Masanari KONDO†, Yutaro KASHIWA†, Yasutaka KAMEI†, and Naoyasu UBAYASHI†

† Kyushu University
E-mail: †{kuramoto,ishimoto,shindo}@posl.ait.kyushu-u.ac.jp,
††{kondo,kashiwa,kamei,ubayashi}@ait.kyushu-u.ac.jp

Abstract　A quick and accurate understanding of the symptoms of a bug can greatly affect the efficiency of debugging. Videos of bug reproductions and images showing the status of bugs are expected to be useful in the debugging process. The purpose of this study is to clarify the actual usage of bug report videos and images in software development sites. In this study, we investigated the impact of including video or images in the issue reports of about 4,000 public repositories in GitHub. As a result, the average time to resolve an issue report increased by 2.9%~17.6% and the average number of comments increased by 76.4%~143% for issue reports that included video and images compared to those that did not. On the other hand, the average time until the first comment decreased by 20.1%~25.4%. We report the details of these results.

Key words　GitHub, Movie, Image, Bug report, Issue

1. はじめに

ソフトウェア開発を行う際の重要な工程の一つとしてデバッグ作業があげられる．デバッグ作業がソフトウェアの開発コストの50%以上を占めるという結果が示されている[1], [2]．この問題を解決するためにバグ修正効率化の研究が盛んに行われている．

バグの症状の迅速で的確な把握は，バグの再現性を高

め，デバッグ作業の効率に大きく影響する点で重要であると考えられる．バグの再現動画及びバグの状況を表す画像は，バグの症状把握に有用であり，これを有効に活用することでデバッグ作業の効率化が期待できる．

本研究の目的は，ソフトウェア開発現場におけるバグ報告動画及び画像の活用実態を明らかにすることである．GitHub で公開されている約 4,000 件のリポジトリを対象に，課題報告（Issue レポート）において動画及び画像の有無による影響を調査した．その調査結果について，報告する．

本稿では，2 節で本研究を行うに至った動機を述べ，3 節で本調査で用いる分析法について述べ，4 節で本研究で用いるデータセットについて述べる．5 節で調査内容と検定結果を述べ，6 節で調査結果と考察を述べる．7 節で妥当性への脅威について述べ，最後に 8 節でまとめと今後の課題について述べる．

2. 動　　機

GitHub では，公開されたリポジトリに対して，そのオーナーだけでなく不特定多数のユーザーが開発に貢献しており，不具合や意見の報告時に Issue 機能が使用されている．2021 年以前は，Issue 作成には文字入力の他に，GIF 動画や画像のみ添付可能であり，動画添付時には一度 GIF 動画に変換する手間があった．現在では，MP4，MOV ファイルを容易に添付できるようになり，今後動画の活用件数は増加すると考えられる．しかし現時点では，動画及び画像の有無による Issue への影響について明らかにされていない．本調査では，解決までの時間や，コメントの回数等取得可能な Issue のパラメータを動画及び画像の有無を考慮して比較する．また，動画及び画像がどのような単語と併用されているかを調査することで，どのような報告内容のときに動画及び画像が用いられているかを調査する．

3. 本調査で用いる分析法

本節では，本調査で用いた検定法及び指標について紹介する．

3.1 検　定　法

本調査では，画像を含む Issue，動画を含む Issue，どちらも含まない Issue の 3 群の比較を行う．差の検定法では一般に，t 検定が用いられるが，3 群のそれぞれの組み合わせに対して 1 回あたり有意水準 α で検定を 3 回繰り返すと、真の有意水準が α よりも大きくなってしまう．誤って有意差があると判断してしまうことにつながる（第一種の過誤）．3 群以上の群間比較では，比較回数に応じて有意水準または有意確率を調節する多重比較が行われる．本調査では，正規性及び等分散性を満たさないデータに対しても検定できる点において，本調査のデータの性質に合致する Steel-Dwass 法を採用する．Steel-Dwass 法は，ウィルコクソンの順位和検定に基づくノンパラメトリックな多重比較法である．ウィルコクソンの順位和検定は，以下の概念による検定法である．データがいずれの群のものであるかに関わらず，全てのデータに対して順位を付ける．さらに順位に重み付けした値を各群ごとに合計し，その順位和を比較する．群間に差がなければ，それぞれの順位和は同程度になると考えられる．これらの間に大きな差があるとすれば，ある群が順位の低いものになり，他の群が順位の高いものになる傾向があると判断する．結果，各データ群は異なる母集団から抽出されたものである，あるいは，統計的に有意な差があると結論づける．従って，帰無仮説は「すべての群の母集団分布は同じ」であり，棄却された時の解釈は「同じ母集団から採られたものではない」となる．ここで注意しなければならない点は，「群 A は群 B よりも有意に大きい」と解釈できるのは，すべての群を通して等分散性が成立するときである．[3] また，順位和検定は順位を用いるため，平均値よりも中央値付近に強く影響を受ける．その他，正規性の検定法には，コルモゴロフ・スミルノフ検定を採用し，等分散性の検定法には，Levene 検定を採用する．

標本化について．比較する各群から同一サイズの標本を抽出し，有意確率を算出する試行を 10 回行う．10 回の平均値を検定で得た有意確率とする．サンプルサイズは、用いる有意水準，検出力，効果量及び用いる検定法によって決まる．有意水準は一般的な 0.05 とする．その他の水準には，やや恣意的に検出力を 0.80，効果量を 0.10 とする．これらと，3 群の対応のない比較の条件のもとで，各群から 323 サンプルとする．サンプルサイズの算出には **G*Power** [4] [5] を利用した．効果量は一般に先行研究からわかっている効果量を用いるが，本調査ではそれが得られなかったために，0.10（効果量小），0.25（効果量中），0.40（効果量大）の **Cohen（1988）**の基準から小程度の効果量を採用した．なお，多重比較法における検出力の考え方についての枠組みは，サンプルサイズの設計という観点では，明確化されていない様である．[6] 効果量，及び，検出力をやや恣意的に決定していることによるサンプルサイズへの影響は，6 節で考察する．

3.2 tf_idf

tf_idf とは，文書中に含まれる単語の重要度を評価する指標である．**tf** とはある語彙の出現頻度であり，**idf** とはある語彙の出現する文書数の逆数を取ったものである．**tf_idf** は **tf** と **idf** を掛け合わせたものである．語

彙 **t** が文書 **d** に出現する回数を **f**，**d** の全単語数（重複を許す）を **T**，**t** が出現する文書数を **n**，全文書数を **N** とすると，**d** 内の **t** に対する **tf_idf** は以下の式で求められる．

$$\mathbf{tf(t,d)=f/T}$$
$$\mathbf{idf(t)=log(N/n)}$$
$$\mathbf{tf\_idf(t,d)=tf(t,d)*idf(t)}$$

tf_idf 値が大きい単語は，その文書において重要な意味を持つと考えられる．Issue 群ごとに，**tf_idf** 値が大きい単語を算出し，どんな内容の Issuen に対して動画及び画像が用いられる傾向があるかを分析するために用いる．

4. データセット

本節では，本研究で用いるデータセットについて紹介する．

4.1 本調査の対象

次の条件を満たすリポジトリを対象とした．

- スターが 10 個以上あること
- 最終更新が 2021 年以降であること
- issue が一つ以上あること

スター数の条件は，本研究に関連する可能性が低いプロジェクト（個人のプロジェクトやサンプルプロジェクトなど，不具合報告が発生しにくいもの）を除外するために設定した．また，最終更新を 2021 年以降に限定することで，より近年の動向を得られると判断した．すべての条件を満たすリポジトリは，289,115 件であった．これらのリポジトリから 4,173 件のリポジトリをランダムに選んだ．4,173 件のリポジトリから，770,656 件の解決済みの Issue を取得した．データ収集期間は 2021 年 11 月から 12 月であり，件数は当時のものである．

4.2 データ取得内容と方法

本調査で用いる Issue の 8 つの調査項目を紹介する．調査項目を表 1 に示す．

Issue の調査項目は，GitHub API である PyGithub（python ライブラリ）を用いて取得した．Num_of_char, Num_of_img, Num_of_mov, $Words$ は直接取得できないため，その取得方法を紹介する．Issue に貼り付けた動画及び画像は URL に変換され，Issue のテキスト（直接取得可能）にマークダウン形式で記述される．URL に変換された一例を次に示す．

https://user-images.githubusercontent.com/XXX.mp4

XXX の部分は，半角英数字，"/"，"-"のみで構成される．
従って，以下の正規表現に適合した文字列の出現回数を，拡張子 Y に応じて Num_of_img, 及び, Num_of_mov とした．

https://user-images.githubusercontent.com/[\w/-]+.Y

表 1　issue の調査項目

調査項目	内容
$Issue\_open\_time$	Issue が解決するまでの時間（日）
$First\_comment\_time$	最初のコメントがつくまでの時間（日）
$Num\_of\_comments$	寄せられたコメント数
Num_of_char	Issue の作成時の文字記述量
Num_of_img	Issue の作成時の画像添付数
Num_of_mov	Issue の作成時の動画添付数
$Words$	Issue の作成時の記述された英単語
$Issue\_created\_at\_year$	Issue が作成された年

$Words$ は，Issue のテキストから正規表現で取得した．また，Issue のテキストの長さを，Num_of_char とした．動画及び画像を含む Issue については，上記 URL は除外する前処理を行った．

我々は事前調査で，$Issue\_open\_time$ が，何らかの要因により負値をとる Issue が混在していることを発見した．このような不正な値を含む Issue を除外するため，30seconds <= $Issue\_open\_time$ <= 1year の条件を設けた．この条件を満たす Issue は，711,160 件（92.23%）であった．

4.3 データの分類

動画及び画像の有無で Issue を次の 3 つのカテゴリに分類した．

Img　Num_of_img が 1 以上の Issue
Mov　Num_of_mov が 1 以上の Issue
$None$　上記どちらにも含まれない Issue

ただし，動画と画像をどちらも含む Issue は Img 及び Mov の両方に属す．各 Issue 群の要素数を表 2 に示す．

表 2　カテゴリー分類・要素数

Img	Mov	$None$
33,079 (4.65%)	3,819 (0.54%)	674,793 (94.81%)

収集したデータの一部抜粋を表 3 に示す．1 つの Issue の情報が表の行に並んでおり，$Issue\_open\_time$ 及び $First\_comment\_time$ の単位は日数である．

表 3　取得した Issue の調査項目の一部抜粋

Issue_created_at_year	Issue_open_time	Num_of_img	Num_of_mov	Num_of_comments	First_comment_time	Num_of_char
2020	6.99861111	0	0	1	6.99861111	39488
2020	41.9594329	1	0	3	17.7784722	950
2020	43.8850579	0	0	2	0.49828704	379
2020	44.0935532	0	0	4	0.91277778	191
2020	0.14934028	0	0	8	0.08077546	1800
2020	59.5670949	2	0	5	0.39472222	999
2020	74.9322569	0	0	0	-	127

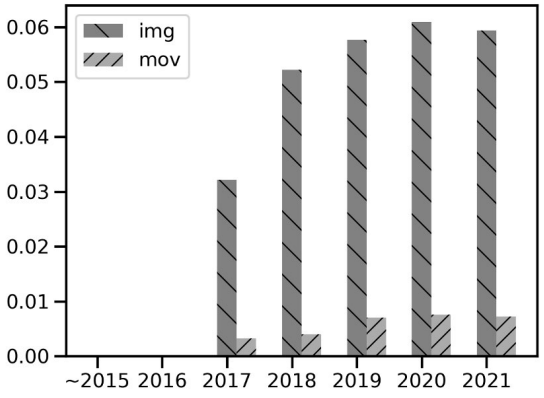


図 1　動画・画像の含有率推移

5. 調　　査

5.1　動画・画像の活用件数の推移

図 1 は, 動画及び画像の活用件数の含有率の年別推移である．2016 年までは，画像の活用は 0.01% 未満であり，動画に関しては 0 件であった．2017 年以降，活用件数は急激に増加し，2021 年には，動画 0.72%，画像 5.93% であった．2017 年以降に急激に増加した要因に関しては，調査中である．

5.2　データの分析

図 2 は, データの特性値を箱ひげ図及び表で表現したものである．ただし, 図 2 の箱図の最大値は, $3rd\,Qu+1.5*IQR$ 以下の最大値である．（IQR：四分位範囲）

正規性について．コルモゴロフ・スミルノフ検定の結果，すべての項目に対して，正規性は有意水準 0.05 で棄却された．

等分散性について．Leneve 検定の結果を表 4 に示す．帰無仮説は「すべての Issue 群間を通して等分散」であり，有意水準は 0.20 とする．

$Num\_of\_comments$ 及び Num_of_char は, 等分散性が有意水準 0.20 で棄却される．$Issue\_open\_time$ 及び $First\_comment\_time$ は，有意確率が有意水準以上であることと，平均分散比が 1 ~ 1.5 程度であることから，等分散性を仮定する．

表 4　等分散性検定・結果

	有意確率	平均分散比
$Issue\_open\_time$	0.378	1.050
$First\_comment\_time$	0.296	1.301
$Num\_of\_comments$	0.001	2.459
Num_of_char	0.073	3.156

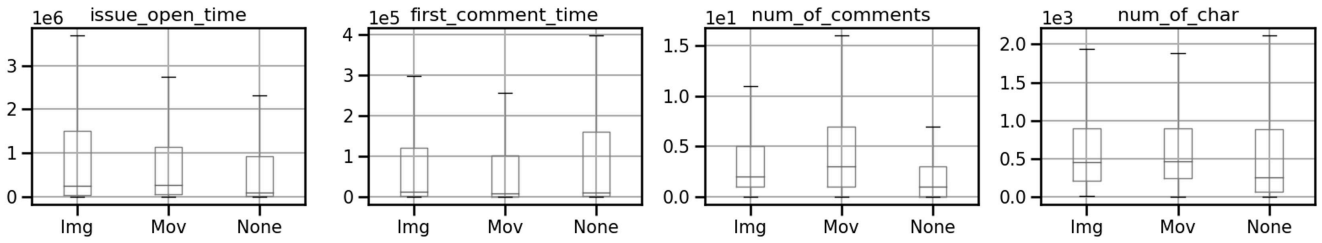
表 5　調査項目ごとの多重比較の結果

調査項目	対	有意確率
$Issue\_open\_time$		
	Img **vs** $None$	0.002
	Mov **vs** $None$	0.021
	Img **vs** Mov	0.381
$First\_comment\_time$		
	Img **vs** $None$	0.764
	Mov **vs** $None$	0.351
	Img **vs** Mov	0.404
$Num\_of\_comments$		
	Img **vs** $None$	0.001
	Mov **vs** $None$	0.001
	Img **vs** Mov	0.211
Num_of_char		
	Img **vs** $None$	0.001
	Mov **vs** $None$	0.001
	Img **vs** Mov	0.599

5.3　統計的検定

ここでは，群間の分布の差が統計的に有意であるかを検定する．Steel-Dwass 法による検定結果を表 5 に示す．ただし，有意水準は 0.05 とする．

Img 及び Mov は，$None$ に比べて，$Issue\_open\_time$ が有意に大きく, $Num\_of\_comments$ 及び Num_of_char については分布に何らかの有意な差があると解釈できる．一方で，$First\_comment\_time$ においては，有意差は認められなかった．また，Img と Mov の比較では，すべての項目で有意差は認められなかった．有意差が認められなかった項目については，有意差の有無は判断を保留する．



	Img	*Mov*	*None*
Mean	24.86	21.77	21.14
Min	30sec	38sec	30sec
1rd Qu	10.2hrs	15.7hrs	2.14hrs
Median	2.96	3.01	1.09
3rd Qu	17.30	13.12	10.80
Max	364.56	361.83	364.99
S.D.	55.64	53.96	53.64

	Img	*Mov*	*None*
Mean	6.83	7.31	9.15
Min	1sec	1sec	1sec
1rd Qu	10.6min	3.38min	10.4min
Median	3.26hrs	2.4hrs	3.0hrs
3rd Qu	1.38	1.18	1.84
Max	364.10	337.61	364.61
S.D.	28.27	32.77	34.39

	Img	*Mov*	*None*
Mean	4.04	5.56	2.29
Min	0	0	0
1rd Qu	1	1	0
Median	2	3	1
3rd Qu	5	7	3
Max	283	120	439
S.D.	6.873	8.710	4.619

	Img	*Mov*	*None*
Mean	995.2	844.8	1350.62
Min	0	0	0
1rd Qu	209	237	59
Median	449	468	251
3rd Qu	901	897	880
Max	260,173	58,887	261,432
S.D.	$4.142E3$	$2.228E3$	$4.811E3$

図 2　Issue 群ごとの調査項目の分布及び代表値

表 6　Issue 群ごとの **tf_idf** 上位 10 単語

	Img	*Mov*	*None*
1	image	when	dependabot
2	img	dropdown	code
3	src	view	file
4	width	package	pullrequest
5	screenshot	issue	version
6	when	python	error
7	error	height	use
8	screen	react	lib
9	version	src	add
10	shot	button	commit

5.4　出現単語分析

Issue のテキストに含まれる単語を **tf_idf** 値に変換し，同じ Issue 群に属すもの同士をマージした結果，**tf_idf** 値が大きい順に 10 個の結果を表 6 に示す．"at"や"it"，"the"などの単語を除外する為，名詞，動詞及び疑問詞のみを対象とした．

6.　結果と考察

画像及び動画を含む Issue は，そうでない Issue に比べて，$Issue\_open\_time$ が平均で，2.9%~17.6% 増加し，$Num\_of\_comments$ が平均で 76.4%~143% 増加した．一方で，$First\_comment\_time$ が平均で 20.1%~25.4% 減少し，Num_of_char が平均で 26.3%~37.5% 減少した．

従って，動画及び画像を用いることには，次の様な効果があると推察できる．

- 状況報告を細かく記述せずに済む
- 開発者の問題への着手を早める
- コメントが得やすくなる

また，$Issue\_open\_time$ 及び $Num\_of\_comments$ が増加していることから，動画及び画像は，比較的難しい問題に対して用いられているのではないかと推察される．

また，Img と Mov の間では，どの項目においても有意な差は認められなかったことから，動画と画像のどちらを用いるかよりも，用いるか用いないかが重要であると推察される．

出現単語分析では，Img では表示に関係する単語が多く，Mov では"dropdown"，"button"及び"react"など動きのある画面に関する単語が多いことを実験的に確かめた．一方で，$None$ では，コーディングに関する単語が多い結果であった．Mov では"when"が 1 番目に，Img でも 6 番目にランクインした．これは，不具合の再現方法に関する記述である可能性が高い．実際，目視調査で，不具合の再現動画が多いことが確認できた．一方で，$None$ では 1 番目に"dependabot"がランクインした．DependaBot は，最新のライブラリを推奨したり更新するボットであり，設定により自動で Issue を生成することもある．Img 及び Mov と $None$ の差は，ボットによる影響を受けている可能性がある．

単語分析の結果から，不具合が画面に関する場合に画像が用いられ，画面の動きに関する場合に動画が用いられ易いと推察できる．

今回，すべての調査項目が正規分布に従わなかった．一般に，正規分布を仮定しない検定法よりも，正規分布を仮定する検定法は検出力に長けていると言われている．従って，データが正規分布に従わない場合，正規分布に近似するため適切な変数変換を行う．時系列データの変換には対数変換が多くの場合有効である．本調査の調査項目の一部は時系列データであり，変換が推奨される場

面であるが，実際に適用させたところ平均値の群間順序関係が変化したことと，順位和検定により順位に変換されることから，対数変換せずに検定を行った．

また，一般にサンプルサイズは多ければ多いほど，母集団の特徴をより正確に抽出できる．ところが検定においては，サンプルサイズが大きくなるにつれて，誤差に過剰に反応し，有意確率が小さくなる傾向がある．すまわち，検定で有意差を得たとしても，母集団分布の有意な差によるものか，サンプルサイズが大きすぎることによるものか判断できないのである．従って，適切なサンプルサイズを前もって決めておくことが重要である．本調査では，事前に効果量の目安が得られなかったために，やや恣意的に決定した．検定結果を見ると，有意差がはっきりと認められた部分と，認められなかった部分が見られた $Issue\_open\_time$, $Num\_of\_comments$ 及び Num_of_char に対しては適切なサンプルサイズであったと推察する．一方で $First\_comment\_time$ に対しては，実際に有意と言える差が無いか，あるいは，サンプルサイズが小さすぎた可能性がある．

7. 妥当性への脅威

内的妥当性．本調査で用いた等分散性の検定の有意水準について，留意しなければならない点がある．多重比較法は等分散性に関して，二群比較よりも鋭敏であり，有意水準 0.20 程度で行わなければ意味をなさないことが示されている．[7] さらには，有意水準 0.50 まで引き上げても，誤った結果を招く場合があると言われている．その場合，$Issue\_open\_time$ 及び $First\_comment\_time$ の等分散性は棄却され，「動画及び画像がある Issue は，そうでないものに比べて $Issue\_open\_time$ が有意に大きい」とした表現は，「動画及び画像がある Issue は，そうでないものに比べて $Issue\_open\_time$ に何らかの有意な差をもたらす」となる．

また，ボットが $None$ に多くあらわれていることが，出現単語分析で明らかになった．bot を取り除く処理をしなければ，分布差が動画及び画像によるものとは断定できない．その点が考慮されていない．

Issue で報告される内容は，不具合に関するものだけではない．事前調査では，新たな機能の提案に関する報告があることも確認した．ただこの場合に関しても，動画及び画像により提案者の意図が伝わりやすくなることは，バグの症状理解が早まることと本質的に同じと考えている．

外的妥当性．本調査の調査対象は，条件に適合したリポジトリのみであり，すべての開発現場で同様の傾向があるとは限らない．また，開発規模等のリポジトリの性質が考慮されていない．

8. まとめと今後の課題

本研究では，ソフトウェア開発現場におけるバグ報告動画及び画像の活用実態を明らかにするため，GitHub で公開されているいくつかのリポジトリを対象に，Issue と動画及び画像の有無の関係を調査した．

スターが 10 個以上のリポジトリに対して Issue 報告を行うのは，一般に開発中級者以上であると考えられる．本調査は中級者以上のバグ報告の実態調査であり，従ってこの試みは，開発初級者がバグ報告する際の参考になると考えられる．

また，今後の課題として，次の 3 つの追加調査を計画している．

- ボットを可能な限り除外すること
- 開発規模の大小で分類することで，実態をより細かく調査すること
- 対象のリポジトリを増やし，より全体の傾向を結果に反映すること

謝　　辞

本研究の一部は XXX の助成を受けた．

文　　献

[1] James S. Collofello and Scott N. Woodfield. Evaluating the effectiveness of reliability-assurance techniques. *Journal of Systems and Software*, Vol. 9, No. 3, pp. 191–195, 1989.

[2] L. Gazzola, D. Micucci, and L. Mariani. Automatic software repair: A survey. *IEEE Transactions on Software Engineering*, Vol. 45, No. 1, pp. 34–67, 2019.

[3] 山本光司. 植物防疫基礎講座 正しい分散分析結果を導くための変数変換法. 第 56 巻, pp. 436–441, 2002-10.

[4] Lang. Albert-Georg Faul. Franz, Erdfelder. Edgar and Buchner. Axel. G*power 3: A flexible statistical power analysis program for the social; behavioral; and biomedical sciences. Vol. 39, pp. 175–191, 2007.

[5] Buchner. Axel Faul. Franz, Erdfelder. Edgar and Lang. Albert-Georg. Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. Vol. 41, pp. 1149–1160, 2009.

[6] 永田靖. 多重比較法の実際. 第 27 巻, pp. 93–108, 1998.

[7] 阿部研自. 多重比較法における不等分散の影響評価. 第 28 巻, pp. 55–78, 1999.